METHODS IN MEDICAL RESEARCH

Volume 6 · J. M. STEELE, *Editor-in-Chief*

3565 ① human genetics
② climatic variables ③ respira
-tory exchange ④ metabolism
⑤ skin temperature ⑥ Sweating
⑦ statistics ⑧ metabolism
cages ⑨

268.	5-5-60.	9 | 11
295.	16|3.	B | 3
——	21.1.67	19/1/67

METHODS IN MEDICAL RESEARCH

VOLUME 1

VAN R. POTTER, *Editor-in-Chief*

Assay of Antibiotics, *Henry Welch, Editor*
Circulation—Blood Flow Measurement, *Harold D. Green, Editor*
Selected Methods in Gastroenterologic Research, *A. C. Ivy, Editor*
Cellular Respiration, *Van R. Potter, Editor*

VOLUME 2

JULIUS H. COMROE, Jr., *Editor-in-Chief*

Methods of Study of Bacterial Viruses, *Mark H. Adams, Editor*
Pulmonary Function Tests, *Julius H. Comroe, Jr., Editor*
Assay of Hormonal Secretions, *Eleanor H. Venning, Editor*

VOLUME 3

RALPH W. GERARD, *Editor-in-Chief*

Genetics of Micro-Organisms, *S. E. Luria, Editor*
Assay of Neurohumors, *J. H. Gaddum, Editor*
Selected Psychomotor Measurement Methods, *Walter R. Miles, Editor*
Methods for Study of Peptide Structure, *Choh Hao Li, Editor*

VOLUME 4

MAURICE B. VISSCHER, *Editor-in-Chief*

Histochemical Staining Methods, *George Gomori, Editor*
Fluid and Electrolyte Distribution, *Louis B. Flexner, Editor*
Studies on Gastrointestinal Pressures, Innervation and Secretions, *J. P. Quigley, Editor*
Tissue Culture Methods, *C. M. Pomerat, Editor*

VOLUME 5

A. C. CORCORAN, *Editor-in-Chief*

Methods for Separation of Complex Mixtures and Higher Molecular Weight Substances, *Lyman C. Craig, Editor*
Methods of Renal Study, *A. C. Corcoran, Editor*
Immunochemical Methods of Determining Homogeneity of Proteins and Polysaccharides, *Melvin Cohn, Editor*

METHODS IN MEDICAL RESEARCH

Volume 6

METHODS IN

Medical Research

Volume 6

J. MURRAY STEELE, *Editor-in-Chief*

SOME METHODS OF STUDYING HUMAN GENETICS, *Antonio Ciocco, Editor*

METHODS IN ENVIRONMENTAL RESEARCH, *Ray G. Daggs, Editor*

STATISTICS IN MEDICAL RESEARCH, *Donald Mainland, Editor*

DESIGN AND CONSTRUCTION OF METABOLISM CAGES, *Arnold Lazarow, Editor*

THE YEAR BOOK PUBLISHERS, INC.
200 EAST ILLINOIS STREET, CHICAGO

PRINTED IN U.S.A

GOVERNING BOARD PREFACE

When we ventured to launch this series of volumes we hoped and believed that its usefulness and popularity would increase as the number of topics covered grew larger, and we are glad to find that this optimistic view had some basis: the appearance of each volume has stimulated demand for its forerunners. This is due, of course, to the enthusiasm and skill with which successive editors, associate editors, and contributors have approached their tasks, so that a high standard has been maintained. It is perhaps unnecessary to explain that editors and topics are tentatively selected some years in advance, and that, since some sections take longer to complete than could be foreseen, these plans have to be flexible enough to permit some reshuffling. The time is probably approaching when we shall have to consider sections planned to supplement and to bring up to date topics dealt with in the earlier volumes. We feel that there will never be any shortage of desirable subjects; and we note with gratification that there seems to be no serious shortage of expert colleagues willing to undertake the onerous tasks of authorship.

The tasks that fall to us, as members of the Governing Board, are relatively light and pleasant. It was inevitable, though to us regrettable, that we should nevertheless fail to keep the original group together and intact; it is on the other hand very pleasant to welcome to the circle such distinguished scientists as Dr. René J. DuBos and Dr. C. N. H. Long, who share with us the planning of forthcoming volumes.

EDITOR'S PREFACE

> "There are men that will make you books, and turn them loose into the world, with as much dispatch as they would do a dish of fritters."
>
> M. DE CERVANTES, *Don Quixote*

THIS BOOK, the Editor wishes to point out, is the antithesis of such a one. Like previous volumes, it has emerged from the labor of many contributors over a long period of time and is not the result of 20 minutes over the frying pan of sizzling ideas.

In the preface to Volume 5 the Governing Board pointed out that there is really no longer need to attempt to explain or to justify this series. Its purpose is thoroughly established and if, as Dr. A. C. Corcoran suggested in his preface to the same volume, these books have found their way into the laboratories rather than into the libraries, that purpose will have been largely accomplished. The way in which each volume has been received makes the need for them self-evident. The Editor's remarks are, therefore, limited to comments on the present volume.

It might be said of the section on Clinical and Climatological Research that there is some overlapping between the chapters on Energy Metabolism and Metabolic Reference Standards, but Dr. Daggs and the Editor agreed that since the two points of view were so appreciably different, a little overlap was healthy. It seems also of interest to point out how some of the other sections dovetail. Dr. Lazarow describes and gives beautiful illustrations of various types of cages and tells one how to build them. When one has built or bought one's cages, Dr. Mainland very conveniently tells one where to put them in order to get the most out of the study.

Then, too, there is the obvious relation of the statistical approach in planning and evaluating the results of an experiment, in Dr. Mainland's section, to that of the somewhat different use of statistical analysis in Dr. Ciocco's section on genetics.

Though the reader may not want to read all of the detailed techniques included in these sections, the general approach to the problem may prove of interest. It is our hope that by having methodology so well compressed between the covers of a single volume much of the labor of looking up methods will be saved for the workers in the respective fields. By the same token, the labor which

went into the preparation of these sections will become apparent.

The Editor must confess that after the members of the Governing Board have planned the volume, the Associate Editors and contributors have done the work and the publisher has expedited matters with great efficiency, little credit is left for the Editor. He has, however, this opportunity to thank all of those who have given so generously of their time and labor to make this volume a reality and to express the hope that it will help make work lighter in the laboratories of those who use it.

—J. Murray Steele.

CONTRIBUTORS AND REVIEWERS

BREWER, N. R., D.V.M., Ph.D.
Central Animal Quarters, University of Chicago, Chicago, Ill.

CARLSON, LOREN D., Ph.D.
Associate Professor of Physiology, University of Washington School of Medicine, Seattle.

CIOCCO, ANTONIO, Sc.D.
Professor and Head, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pa.

CRUMP, LEE S., Ph.D.
Head of the Statistical Section, Atomic Energy Project, University of Rochester, Rochester, N.Y.

DAGGS, RAY G., Ph.D.
Director of Research, Army Medical Research Laboratory, Fort Knox, Ky.

DILL, DAVID B., Ph.D.
Scientific Director, Medical Division, Army Chemical Center, Md.

GAUNT, ROBERT, Ph.D.
Ciba Pharmaceutical Products, Inc., Summit, N. J.

HAMMOND, E. CUYLER, Sc.D.
Professor of Biometry, Graduate School, Yale University; Director, Statistical Research Section, American Cancer Society, New York.

HANSARD, SAM L., Ph.D.
Senior Scientist, University of Tennessee Agricultural Experimental Station, Knoxville.

HARDY, JAMES D., Ph.D.
Associate Professor of Physiology and Biophysics, Cornell University Medical College; Associate Director, Russell Sage Institute of Pathology, New York Hospital, New York.

HENSCHEL, AUSTIN, Ph.D.
Director of Research, Quartermaster Climatic Research Laboratory, Lawrence, Mass.

HERRERA, LEE, B.S. (Pub. Health)
Instructor in Medical Statistics, New York University College of Medicine, New York.

INGLE, DWIGHT J., Ph.D.
Research Laboratory, The Upjohn Company, Kalamazoo, Mich.

LAZAROW, ARNOLD, M.D.
Associate Professor, Department of Anatomy, Western Reserve University, Cleveland.

LEVENE, HOWARD, Ph.D.
Associate Professor of Mathematical Statistics, Institute for the Study of Human Welfare, Columbia University, New York.

LI, C. C., Ph.D.
Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pa.

LUFT, ULRICH C., M.D.
Research Physiologist, U.S. Air Force School of Aviation Medicine, Randolph Air Force Base, Tex.

MAINLAND, DONALD, M.B., Ch.B., D.Sc., F.R.S.Edin., F.R.S.Can.
Professor and Chairman, Department of Medical Statistics, New York University College of Medicine, New York.

MILLER, A. T., Jr., Ph.D., M.D.
Professor of Physiology, University of North Carolina Medical School, Chapel Hill, N. C.

MOLNAR, GEORGE W., Ph.D.
Research Physiologist, Army Medical Research Laboratory, Fort Knox, Ky.

PITTS, GROVER C., Ph.D.
Assistant Professor of Physiology, University of Virginia School of Medicine, Charlottesville, Va.

PRATT, RICHARD L., B.A.
Meteorologist, Quartermaster Climatic Research Laboratory, Lawrence, Mass.

ROBINSON, ALINE H., A.B.
Indiana University Medical School, Bloomington.

ROBINSON, SID, Ph.D.
Professor of Physiology, Indiana University Medical School, Bloomington.

STEELE, J. MURRAY, M.D.
Professor of Medicine, New York University College of Medicine, New York.

STOLL, ALICE M., A.M.
Teaching and Research Assistant, Cornell University Medical College, New York, N.Y.

WEDGWOOD, RALPH J., M.D.
Captain, M. C., U.S. Army, Quartermaster Climatic Research Laboratory, Lawrence, Mass.

YOUNG, ALLAN C., Ph.D.
Research Associate in Physiology, University of Washington School of Medicine, Seattle.

TABLE OF CONTENTS

SECTION I

Some Methods of Studying Human Genetics

ASSOCIATE EDITOR—*Antonio Ciocco*

INTRODUCTION

GENES ARE THOSE mysterious entities that hold the answer to all human traits—if one can read their evidence accurately. The various methods of studying genes become as important and dynamic as the study of the genes themselves. The purpose in this section is to explore some methods used in the study of human genetics.

Research in human genetics involves the application of the same fundamental concepts which characterize genetic studies on plants and animals. However, there are some basic difficulties in studying human genetics which arise from the eternal problem that observations on experiments conducted by the subjects themselves must replace observations on experimentation designed by the investigator or, simply, man studying man.

In studies in human genetics the first step is, of course, the collection of pedigree data to provide information regarding the incidence of the condition under study. This condition may be a disease or a physiologic or psychologic trait in the familial aggregate. Analyses of the data should provide definite impressions concerning the genetic constitution of the family and the manner in which the observed trait is presumably transmitted from generation to generation. Both these steps, collection of data and analysis, require thorough knowledge of statistical logic and of analytical techniques. In addition, knowledge of biology and of the mechanisms of cellular reproduction are prerequisite to correct analysis. In this section, Dr. Li limits himself to exploration of certain methods

which have, in recent years, been proved to be sufficiently simple for general use in the problems encountered in the field of clinical medicine.

The physician's interest in genetics is aroused by the patient who possesses a certain condition or disease which is "familial." Therefore Dr. Li has limited his description to the analytical techniques of tracing pedigrees through a propositus—the case seen in the clinic or physician's office. To illustrate the analytical methods, examples have been chosen concerning the genetic nature of the following conditions: human albinism, congenital absence of maxillary lateral incisor, sickle cell anemia, allergic asthma, diabetes mellitus, peptic ulcer and mammary cancer.

Three aspects of methodology are covered. The first deals with methods of estimating the proportion of offspring with certain traits born from matings in which parents do not manifest these traits: Segregation of Recessive Offspring. The second deals with the increased knowledge of the genetic nature of several human diseases resulting from improved precision in diagnosis of disease and statistical analysis: The Severity of Abnormality. Methods of establishing the genetic role is the third aspect of methodology discussed. Several recent examples are given to illustrate the current method of approach.

A detailed account of methods of detecting and measuring genetic linkage in human populations is beyond the scope of this section. However, Dr. Li believes it necessary to include a brief discussion of this subject in the closing paragraphs.

As we are dealing with the methods of studying human genetics, the references listed at the end of the section give only examples of the available literature on methods. Readers who are interested in the heredity of particular diseases will have no difficulty in finding pertinent references in their field of interest.

The most conspicuous omission of methods in current use is the series of methods concerned with populational gene-frequency analysis. These have been summarized by Hogben (11), Li (15) and Dahlberg (1). Some excellent papers by Snyder (23, and the references listed there) should also be consulted by those interested. As to the general principles of human genetics, Stern's book (24) is an indispensable reference. Those who are interested in the broad aspect of human genetics may find valuable references given by Strandskov (25) and Muller (18, 19).

—Antonio Ciocco.

I. SEGREGATION OF RECESSIVE OFFSPRING

C. C. LI, *University of Pittsburgh*

METHODS OF ESTIMATING the proportion of offspring possessing traits not apparent in their parents are highly important in studies in human genetics. The indication would be that the factors for their genetic transmission (genes) are present in the parents, but recessive (recessive genes). Therefore, an accurate estimate of the true proportion of recessive offspring from a certain type of mating becomes the first step toward elucidating the genetic nature of the trait. For the sake of concreteness and brevity, we shall refer to individuals bearing certain traits as "affected" and to those without the trait as "normal."

When a trait is absent in both parents but present in their offspring, there is a strong suggestion that the trait is due to the homozygous condition of 1 or more pairs of recessive genes. Let us consider the simplest case, in which the trait is caused by only 1 pair of genes. Then the genotype of the "affected" individual is *aa*, while that of the nonaffected or "normal" individual is either *AA* or *Aa*—the gene *A* being dominant to its allele *a*.

It is evident that the *aa* or "affected" individual can be produced from 3 types of parental combinations: $Aa \times Aa$, $Aa \times aa$ and $aa \times aa$. If we assume that the recessive gene *a* is rare so that there are relatively few *aa* individuals in the general population, it follows that heterozygous individuals would be much more numerous than the homozygous recessives. Since there are few *aa* individuals, the $aa \times aa$ type of mating will be extremely rare, and the $Aa \times aa$ type of mating will also be uncommon. Consequently, most of the "affected" individuals will be the progeny of the matings $Aa \times Aa$ in which both parents are "normal." It is this type of family that we shall consider in the following paragraphs.

It is important to realize that the $Aa \times Aa$ type of mating cannot be distinguished from the other types of mating involving 2 "normal" parents ($AA \times AA$ or $AA \times Aa$) unless they produce at least 1 "affected" child among their offspring. The "affected" child then serves to indicate that both parents are heterozygous. (Matings of "normal" parents when either is of the *AA* genotype could not produce an *aa* child.)

As an example of a simple recessive trait, rare in the general population, let us consider human albinism—the absence of pig-

ment in skin, hair and iris. This trait is found in about 1 in 20,000 individuals in Europe. According to the Hardy-Weinberg law for random mating populations (Stern (24), chap. 10; Li (15), chap. 2), the frequency of the albino gene in the general population would be approximately 1/140, being the square root of the proportion of albino individuals in the population. Consequently, the proportion of heterozygous individuals (apparently normal, but carriers, Aa) in the general population is approximately 1/70 or 1.4%. We have seen that only certain types of matings can produce an "affected" child. Now it should be clear that almost all of the albino individuals are produced by $Aa \times Aa$ matings because $Aa \times aa$ and $aa \times aa$ are much too rare. The only way we can distinguish $Aa \times Aa$ from $AA \times AA$ or $AA \times Aa$ is, however, by the presence of at least 1 albino among the offspring of the particular mating.

This method of determining the parental genotypes by their offspring is quite similar to the concept of "progeny-test" employed by plant and animal breeders. Unfortunately for the study of human genetics, in many cases heterozygous individuals cannot be distinguished from the homozygous dominant individuals without a "progeny-test." Even more unfortunate is the fact that the "progeny-test" in man is not at all efficient because of the small number of children in a family. Since the $Aa \times Aa$ matings can only be identified by their having at least 1 "affected" child, those matings that fail to produce any affected children would not be observed.

The main purpose of the methods to be discussed here is to overcome the bias caused by the omission of some of the $Aa \times Aa$ families which has distorted the classic mendelian ratio. The probability that an $Aa \times Aa$ mating would produce an aa child is $1/4$. If Aa individuals could be distinguished from AA without the "progeny-test" so that we could select a number of $Aa \times Aa$ unions, the total offspring of such matings would consist of 75% "normal" children and 25% "affected."

Suppose that these matings ($Aa \times Aa$) produce 4 children. The probability of all of the children being "normal" is $(3/4)^4$ or 81/256. This fraction of $Aa \times Aa$ families could not be identified and so would be omitted from our observation. The rest of the families, $1 - (3/4)^4$ or 175/256, would have at least 1 "affected" child and thus could be identified. Observe that the total progeny of the 256 families will consist of $1/4$ "affected" persons, but those of the selected 175 families with at least 1 "affected" child will have a much higher proportion of recessives because of the selectivity. Even for larger sibships, where there are 5 chil-

dren, for example, there still will be $(3/4)^5$ or 243/1024 of the $Aa \times Aa$ families unidentified.

The situation is quite similar to Mendel's peas in this respect. Of the total number of peas of the F_2 seeds (derived from self-pollinating F_1 plants), $^3/_4$ are round and smooth and $^1/_4$ are cuboid and wrinkled. But examination of the peas in each single pod, which usually number 4–6, would not always verify the 3:1 ratio. Specifically, among the pods with 4 peas each, 81/256 will contain all round smooth peas and no wrinkled ones. Discarding these 81 pods and counting the peas in the remaining 175 pods, the proportion of wrinkled peas will be much higher than $^1/_4$.

Our problem, then, is how to obtain the correct proportion of recessives in human families when the identifiable families consist of only a part of the real whole.

SOME PROPERTIES OF THE BINOMIAL DISTRIBUTION

Let us consider certain well known properties of binomial distribution which will provide an approach to the solution of this problem.

For the first example, let p be the probability of "success of an event" in a single trial, and $q = (1 - p)$ be the probability of its failure. If we have s independent trials, then the number of successes in the s trials will be distributed according to the expansion of the binomial $(q + p)^s$. Let r be the actual number of successes among the s trials, $P(r)$ the probability of having r successes in s independent trials, then the probability distribution for the various possible results will be

$$\begin{array}{lll} r: & 0, \quad 1, \quad \ldots, & r, \quad \ldots, s \\ P(r): & q^s, spq^{s-1}, \ldots, & \binom{s}{r} p^r q^{s-r}, \ldots, p^s \end{array} \tag{1}$$

The sum of $P(r)$ from $r = 0$ to $r = s$ is of course 1. If we multiply each $P(r)$ by its corresponding r, the resulting series of quantities, $rP(r)$, will be proportional to the terms of the expansion of a binomial of a lower degree, that is, $(q + p)^{s-1}$. This is easily seen when the common factor sp is removed from each term of the series $rP(r)$. Then the series can be expressed as

$$sp\left[0, q^{s-1}, (s-1)pq^{s-2}, \ldots \binom{s-1}{r-1} p^{r-1} q^{s-r}, \ldots p^{s-1}\right] \tag{2}$$

As a numerical illustration, consider the binomial expansion in which the probability of success $p = 1/4$. If we have $s = 5$ independent trials, the distribution of the number of successes, r, will

be given by the terms of $(3/4 + 1/4)^5$. Thus: writing the common denominator of the fractions $P(r)$ as the "sum" for convenience

r:	0,	1,	2,	3,	4,	5	Sum
$P(r)$:	243,	405,	270,	90,	15,	1	/ 1024
$rP(r)$:	0,	405,	540,	270,	60,	5	/ 1280

Note that the last row is proportional to the terms of $(^3/_4 + ^1/_4)^4$:

$r' = r - 1$:	0,	1,	2,	3,	4	Sum
$P(r') = P(r - 1)$:	81,	108,	54,	12,	1	/ 256

thus

$$\frac{405}{1280} = \frac{81}{256}, \quad \frac{540}{1280} = \frac{108}{256}, \text{ etc.}$$

This indicates the important fact that a binomial distribution of s degree can be reduced to a binomial distribution of $s - 1$ degree by the simple operation of multiplying each term of the original distribution by its corresponding value of r. It may be noted that in this operation, it is not necessary to know the value of $P(0)$ corresponding to $r = 0$ since this term will drop out in the reduced distribution.

Let us now consider another question. Suppose that we are given a *complete* binomial series, how would we proceed to find the value of p which has given rise to the distribution? This question can be solved in various ways. The simplest method is to derive p from the mean number of successes ($\bar{r}$) per s independent trials. From the property of binomial distribution we have just described, it follows from formula (2) that

$$\bar{r} = \sum_{r=0}^{s} rP(r) = sp(q + p)^{s-1} = sp \tag{3}$$

From this formula the value of p can be obtained immediately. For example, if we are given the series 243, 405, 270, 90, 15, 1 corresponding to $r = 0, 1, 2, 3, 4, 5$ where $s = 5$, then the mean value of r is found to be

$$\bar{r} = \frac{1280}{1024} = sp = 5p,$$

Therefore

$$p = \frac{1280}{1024} \times \frac{1}{5} = \frac{1}{4}$$

If the given series lacks the first term corresponding to $r = 0$ so that it consists only of the numbers 405, 270, etc., corresponding

to $r = 1, 2$, etc., we can still find the value of p by the simple operation of first reducing it to a *complete* series of a lower degree, and then following the procedure given above. In the case of the previous example, the *truncated* distribution for $s = 5$ will be reduced to a complete distribution for $s' = s - 1 = 4$ and the mean value of

$$\bar{r}' = \sum_{r=0}^{s'} r'P(r') = \frac{256}{256} = s'p = 4p$$

Therefore, $p = 1/4$, as before.

The methods to be described—proband-method, sib-method and maximum likelihood estimate—involve the same principles outlined here, either by direct application or by some modification.

PROBABILITY OF DETECTING A SIBSHIP

Our problem, as previously stated, is how to obtain the correct proportion of recessives in human families where the identifiable families consist of only a part of the real whole. We now approach this problem by utilizing the properties of the binomial distribution already discussed.

Let us consider the families with s children, s being the "size" of a sibship. Let r denote the number of "affected" children in a sibship of s members where p is the probability that a child should be "affected." Thus, $q = (1 - p)$ is the probability of a child's being "normal." The various kinds of sibships ($r = 0, 1, 2, \ldots, s$) will be distributed according to the binomial expansion of $(q + p)^s$ (1). But there will be q^s of these families without any affected child. Therefore identifiable sibships consist of only $1 - q^s$ of the total sibships in which both parents are heterozygous. The distribution of these identifiable sibships is therefore a truncated binomial series lacking the first term

$$\begin{array}{lccccc} r: & 1, & & 2, & \ldots, & s \\ P(r): & \dfrac{spq^{s-1}}{1-q^s}, & & \dfrac{\binom{s}{2}p^2q^{s-2}}{1-q}, & \ldots, & \dfrac{p^s}{1-q^s} \end{array} \qquad (4)$$

so that the sum of $P(r)$ from $r = 1$ to $r = s$ is unity. This can be considered as the "universe" of sibships of size s from which we observe a number of families—the sample. Let a_{rs} be the observed number of sibships of size s with r "affected" members; thus

$$\sum_{r=1}^{r=s} a_{rs} = n.$$

is the total observed number of sibships of size s. Then our problem is to estimate the value of p from the observed series a_{rs}.

$$\begin{array}{l|l} r\colon\ 1,\ 2,\ \ldots,\ r,\ \ldots,\ s & \text{Sum} \\ a\colon\ a_{1s},\ a_{2s},\ \ldots,\ a_{rs},\ \ldots,\ a_{ss} & n_s = \Sigma a_{rs} \end{array} \tag{5}$$

It should be clear that the estimate of p depends on the distribution of a_{rs} in the total of n_s families. In turn, the number of families (the proportional partition of n_s into its parts), a_{rs}, depends upon the chances of detecting a sibship with its various values of r. Hence we must seek to examine what effects the different values of the probability of detecting a certain sibship will have on the relative values of a_{rs}.

First, consider the simplest case where each sibship is equally likely to be detected, regardless of the value of r (number of affected members present in the sibship) provided, of course, r equals at least 1. In this case, if we have a large number of sibships of size s, it is obvious that the observed numbers a_{rs} will be proportional to the probability distribution as given by formula (4).

However, the assumption that a sibship with $r = 4$ is as likely to be detected as one with $r = 1$ is valid only when our search for affected offspring in a population is complete so that every family with any "affected" children at all will be detected and noted. If this could be accomplished, the observed a_{rs} should be distributed according to formula (4), obviously, because the whole universe has been ascertained. To state it in a more rigorous manner, let π be the probability of having a single affected *individual* detected; if $\pi \to 1$ so that almost every affected child in a population has been ascertained (many of them belonging to the same family, of course) the distribution of a_{rs} will be of the form (4). To be more realistic, we are fully aware that genetic investigations are based on the collection of data on persons visiting a clinic or a physician's office, or are selected in some other manner, so that π will be much smaller than unity.

Let us, therefore, consider our problem from another angle. The probability that an affected *individual* will not be detected is $1 - \pi$. Hence, the probability that a *sibship* containing r "affected" will escape detection is $(1 - \pi)^r$, so that the probability that it will be detected is $1 - (1 - \pi)^r$. It follows that the larger the value of r in a sibship, the more frequently it will be detected. Thus, if the probability of detecting an "affected" individual is $1/3 = \pi$, the sibships with 2 affected members will have a 55% chance of being detected while those with 4 affected members will be detected in 80% of the cases. In particular, if $\pi \to 0$ (very small, as when we have only a small collection of "affected" per-

sons who just happen to come to our notice), then the probability of detecting a sibship with r affected becomes (7)

$$1 - (1 - \pi)^r \rightarrow \pi r$$

That is, the probability of detecting a sibship is almost exactly proportional to r, the number of "affected" in that sibship. Thus the sibships with 4 "affected" persons will be detected twice as frequently as those with 2 "affected" persons, and 4 times as frequently as those with only 1 "affected" member. Consequently, the distribution of a_{rs} will be that of (4) multiplied by the corresponding value of r for each term. In other words, the form of the distribution of families will be a complete binomial series but of a lower degree as given by formula (2).

To summarize, when the value of π is close to unity, the a_{rs} series forms a truncated binomial series of degree s; when π is close to 0, the a_{rs} series forms a complete binomial series of degree $(s - 1)$. In practice, however, it is difficult to obtain exact knowledge concerning the magnitude of π, the detection-probability. If the data were collected in such a way that the number of times a sibship has been independently detected is recorded, the value of the detection-probability can, according to Fisher (4), be estimated. For intermediate values of π, the distribution of a_{rs} is a function of π as well as of s, r and p. Furthermore, π may vary from family to family and may be a complicated function of family size and number of affected children. In the following sections we shall only deal with the two cases when $\pi \rightarrow 1$ and $\pi \rightarrow 0$.

SIBSHIPS OF A GIVEN SIZE

For the first illustration, let us consider the methods of estimating p from sibships of the same size (s remaining constant). Our observed data consist of the simple series of numbers (5). In all cases that follow we assume that ascertainment is nearly complete ($\pi \rightarrow 1$) except where there is a statement to the contrary.

(i) *Proband method.*—Since ascertainment is nearly complete, the observed a_{rs} series should be proportional to terms of (4). Its theoretical mean value of r is $\Sigma rP(r) = sp/(1 - q^s)$ while our observed mean is $\bar{r} = \Sigma r a_{rs}/n_s = r_s/n_s$, where $r_s = \Sigma r a_{rs}$, the total number of "affected" members in the n_s sibships of size s. Hence our estimate of p is obtained by solving the following equation

$$\bar{r} = \frac{sp}{1 - q^s} \tag{6}$$

or, since $r_s = \bar{r}n_s$

$$\frac{r_s}{p} = \frac{sn_s}{1 - q^s} \tag{6'}$$

Note that equation (6) is similar to (3) except for the correction factor $1 - q^s$. As a numerical example, suppose we have the following observed data which consist of 781 sibships of 5 members each:

r:	1,	2,	3,	4,	5	sum
a_{rs}:	405,	270,	90,	15,	1 /	781 = n_s
ra_{rs}:	405,	540,	270,	60,	5 /	1280 = r_s

The observed mean number of affected per sibship is $\bar{r} = 1280/781$. Our estimation equations are then

$$(6): \frac{1280}{781} = \frac{5p}{1 - q^5}, \quad (6'): \frac{1280}{p} = \frac{5(781)}{1 - q^5}$$

Solving, it will be found that $p = 1/4$ which is the correct solution.

(ii) *Weinberg's sib-method.*—Alternatively, we may reduce our observed data by multiplying a_{rs} by its corresponding value of r so that the resulting series of numbers will be proportional to a complete binomial series of degree $(s - 1) = s' = 4$, and then our formula (3) would be directly applicable provided we replace r by $(r - 1) = r'$ also. After this operation our original observed data will take on the following form

$(r - 1) = r'$:	0,	1,	2,	3,	4	sum
$(ra_{rs}) = a'_{rs}$:	405,	540,	270,	60,	5 /	1280

It will be found that the mean value of r' (not r) with corresponding frequencies 405, 540, etc., is 1280/1280 = 1. Formula (3) gives us the estimation equation

$$\bar{r}' = s'p, \text{ i.e., } 1 = 4p, \quad \therefore p = 1/4$$

The latter equation is much easier to solve than (6) or (6').

When the detection-probability is small ($\pi \to 0$), the original observed numbers of sibships, a_{rs}, would form a complete binomial series of degree $(s - 1)$ without the reduction operation. Replacing r by $(r - 1)$ and s by $(s - 1)$, the calculation of p would be a direct application of formula (3).

(iii) *Maximum likelihood estimate.*—As another choice, we may employ the method of maximum likelihood to estimate the value of p. Since the observed number of sibships of size s containing r affected members is a_{rs}, and the probability of observing 1 such sibship is $\binom{s}{r} p^r q^{s-r}/(1 - q^s)$, the logarithm of the probability of observing such a compound event $(a_{1s}, a_{2s}, \ldots, a_{ss})$ is

$$L = \sum_r a_{rs} \log \left[\binom{s}{r} p^r q^{s-r}/(1 - q^s) \right] \tag{7}$$

It is unnecessary to go through the details of the algebra involved here. Suffice it to say that the estimation of p is obtained by solving the equation $dL/dp = 0$ at which point the value of L will be a maximum. The happy result is that the equation $dL/dp = 0$, upon simplification, reduces to our previous estimation equation (6′). In other words, the maximum likelihood estimate is equivalent to that based upon the mean (or total) number of affected members. Therefore, the value of p as determined by (6′) is the most efficient estimate of the true proportion of affected among the offspring of heterozygous parents.

COMBINING SIBSHIPS OF VARYING SIZES

In any collection of family pedigrees we would naturally have sibships of all sizes. It is required to estimate the value of p from

TABLE 1

Size of Sibships, s	No. of Affected in a Sibship, r: 1	2	3	4	5	...	No. of Sibships of Size s, n_s	Total No. of Affected in n_s Sibships, r_s	Total Offspring in n_s Sibships, t_s
2	—	—							
3	—	—	—	.					
4	—	—	a_{rs}	—			$\sum_r a_{rs}$	$\sum_r r a_{rs}$	sn_s
5	—	—	—	—	—				
6	—	—	—	—	—	—			
.									
	Grand total						N	R	T

pooled data on sibships of various sizes. Now, the observed data would assume the form of Table 1 in which the meaning of the symbols is self-evident.

In dealing with the whole body of data, we make no restrictions or assumptions as to the relative frequency of sibships of varying size. The data may include, for example, any number of sibships of size 3 in conjunction with any number of sibships of size 4, etc. The relative frequency of the various values of n_s is totally irrelevant to the method of estimating p. Hence, each row of Table 1 (with a specified value of s) may be regarded as an independent set of observations. To estimate p based on the whole table is a process of estimating the common value of p for the various rows from the combined information of several independent samples.

The sibships consisting of only 1 child ($s = 1$) have been excluded from the analysis because an "only child" must necessarily be "affected" to be included in our data. Then the mean number of affected in such sibships is always unity and $r_1 = n_1$. Substituting these values in (6) or (6′), we see that both sides of the equations are always equal whatever the value of p. In other words, such sibships provide us no information concerning the value of p.

I. First, consider the case where the ascertainment of "affected" in a population is nearly complete (the detection probability $\pi \rightarrow 1$). Since the rows of Table 1 are independent samples, the L function of the whole table is simply the sum of the separate L's for each s as given by (7). Hence dL/dp is the sum of the separate derivatives for each s. Ignoring the algebraic details, we

TABLE 2.—Observed (Hypothetical) Data Assuming Nearly Complete Ascertainment ($\pi \rightarrow 1$)

s	r				n_s	r_s	t_s
	1	2	3	4			
2	72	12	—	—	84	96	168
3	81	27	3	—	111	144	333
4	108	54	12	1	175	256	700
						496	

state that the condition $dL/dp = 0$ for the entire table, upon simplification, leads to equation (6′) summed over all values of s; that is

$$\sum_s \left(\frac{r_s}{p}\right) = \sum_s \left(\frac{sn_s}{1 - q^s}\right)$$

Writing $\Sigma r_s = R$, the grand total number of affected members in all sibships, and $sn_s = t_s$, the total number of offspring among the n_s sibships of size s, as indicated in Table 1, the above estimation equation assumes the form (Haldane (6, 7))

$$\frac{R}{p} = \sum_s \left(\frac{t_s}{1 - q^s}\right) \tag{8}$$

This is a widely used formula in human genetics. The value of p as determined by this equation is a sort of pooled average of the separate p's that might have been estimated from each row of Table 1. A hypothetical set of data is given in Table 2 to illustrate the arithmetic procedure of this method. Our estimation equation (8) gives

$$\frac{496}{p} = \frac{168}{1-q^2} + \frac{333}{1-q^3} + \frac{700}{1-q^4}$$

This equation is usually solved by iteration, using an initial trial value of p that the investigator thinks to be near its true value. In this hypothetical case it is seen that $p = 1/4$ and $q = 3/4$ is the correct solution. A similar example with a different value of p for a dominant trait is given later.

TABLE 3.—Observed (Hypothetical) Data Assuming $\pi \to 0$

s \ r	1	2	3	4	n_s	r_s	t_s
2	72	24	—	—	96	120	192
3	36	21	4	—	64	96	192
4	27	27	9	1	64	112	256
					224	328	640

By (9), $p = \dfrac{328 - 224}{640 - 224} = \dfrac{104}{416} = \dfrac{1}{4}$

s' \ r'	0	1	2	3	n_s	r_s'	t_s'
1	72	24	—	—	96	24	96
2	36	24	4	—	64	32	128
3	27	27	9	1	64	48	192
					224	104	416

$$p = \frac{R'}{T'} = \frac{104}{416} = \frac{1}{4}$$

II. Next, let us consider the case where the detection-probability is small ($\pi \to 0$) so that the probability of recording a sibship is proportional to the number of affected persons in that sibship. Then the probability of detecting a sibship of size s containing r affected persons is $\binom{s-1}{r-1} p^{r-1} q^{s-r}$ according to formula (2), while the observed number of such sibships is a_{rs}. The L function for a particular value of s is similar to (7) except for replacement of the expression within the square brackets by the present probability. A similar procedure of combining the sibships of various sizes to estimate the common value of p leads to the following equation:

$$\frac{R - N}{p} = \frac{T - R}{q}, \quad \text{or } p = \frac{R - N}{T - N} \tag{9}$$

This is another familiar formula in human genetics, first arrived

at by Weinberg (26) through a different method. This formula simply says that p is estimated by the proportion of total recessives among total offspring with the total number of sibships subtracted from each. An example of this method is given in the upper half of Table 3.

Alternatively, if we replace the actual r by $r' = (r - 1)$ and the actual s by $s' = (s - 1)$ in the data and then calculate the corresponding values r_s' and t_s', our estimate of p would be given directly by the proportion R'/T'. This value is the same as (9) because from each sibship we subtracted 1 recessive and treated the actual size of sibship as 1 less. This alternative procedure is illustrated in the lower half of Table 3.

THE VARIANCE OF AN ESTIMATE

If the estimates are obtained by the method of maximum likelihood, as those given in the previous pages, the sampling variance of p is $V(p) = 1/I$, where $I = -d^2L/dp^2$ and L is the likelihood function. In the following, the expressions for I are given without going into the details of algebraic manipulation (7). For the first case where $\pi \rightarrow 1$

$$I = \frac{1}{pq}\sum_s \frac{(1 - q^s - spq^{s-1})t_s}{(1 - q^s)^2} \tag{8V}$$

the reciprocal of which is the variance of p as estimated by (8). For the second case, where $\pi \rightarrow 0$, we have

$$I = \frac{T - N}{pq} = \frac{(T - N)^3}{(T - R)(R - N)} \tag{9V}$$

The tabulations given by Finney (3) will be helpful to calculate the numerical values of (8V).

Having known the magnitude of $V(p)$, we will briefly discuss the problem of testing the significance of the deviation of the estimated p from a certain theoretical value (1/4, say). As noted before, the value of π is hardly known at all in many practical cases. Usually, the value of p as determined by (8) is higher than the theoretical 1/4 for single factor recessives, while that determined by (9) is lower than 1/4. Let p_1 and p_0 be the estimated values of p by (8) and (9) corresponding to the assumptions that π is nearly 1 and 0, respectively, and σ_{p_1} and σ_{p_0} be the standard errors of p_1 and p_0, respectively. Haldane (7) suggested the following criteria as a test of significance of the difference between the estimated and the theoretical value of p

$$p_1 + 2\sigma_{p1} > 0.25 > p_0 - 2\sigma_{p0}$$

We reject the hypothesis that $p = .25$ only when its estimated value is so large or so small that it falls outside the above limits. Unless we have more accurate knowledge of the value of π, it seems that this procedure is the best at our disposal for practical applications.

Although the foregoing methods originated from the problem of estimating the proportion of recessives among the offspring of $Aa \times Aa$ matings, it is evident that they may be applied equally well when 1 parent is normal and the other "affected," i.e., the matings of the type $Aa \times aa$, where the probability, p, of having an "affected" or "normal" child is in each case $^1/_2$. As a matter of fact, it makes no difference in this case whether the trait in question is dominant or recessive.

To sum up the foregoing points, let us present an example in which the character is dominant. A simple dominant character is the congenital absence of the permanent maxillary lateral incisor teeth. In an investigation of the children attending county council schools in the western part of Middlesex (16), the following data were reported as in Table 4: let r be the number of "affected"

TABLE 4

s \ r	1	2	3	n_s	r_s	t_s
2	11	3	—	14	17	28
3	2	7	0	9	16	27
4	0	1	1	2	5	8
5	0	1	0	1	2	5
				26	40	68

(dominant) members in a sibship. Since all the children were examined, equation (8) may be applied for estimating the value of p. Thus

$$\frac{40}{p} = \frac{28}{1-q^2} + \frac{27}{1-q^3} + \frac{8}{1-q^4} + \frac{5}{1-q^5}$$

Solving, we obtain $q = 0.524$ and $p = 0.476$. Its variance may be found by formula (8V) yielding $\sigma_p = 0.070$, approximately. The theoretical value of p being $^1/_2$, we see that the proportion of "affected" children among sibs does not differ significantly from its theoretical value on the assumption that the trait is due to a dominant gene.

MODIFICATIONS OF THE METHOD

In the collection of pedigrees, it sometimes happens that families with only 1 "affected" child are not accurately detected and recorded. The tendency to dismiss such cases as "sporadic" would lead to a shortage of sibships in which $r = 1$. Conversely, if the trait in question could be produced by nongenetic causes, the inclusion of all families with 1 "affected" child would lead to an excess of such families. Therefore, when there is reason to doubt the accuracy of the number of sibships with only 1 affected member, it is advisable to exclude them from the analysis, i.e., omit the column under $r = 1$ in Table 1. The probabilities of observing a sibship of size s containing $r(\geq 2)$ abnormals have to be adjusted accordingly.

For the case where $\pi \to 1$, this probability becomes

$$\binom{s}{r} p^r q^{s-r}/(1 - q^s - spq^{s-1})$$

whose denominator is unity minus the first *two* terms of the binomial series of degree s. Proceeding in exactly the same manner as before, we obtain the following estimation equation

$$\frac{R}{p} = \sum_s \frac{(1 - q^{s-1})l_s}{1 - q^s - spq^{s-1}}$$

If $\pi \to 0$, the probability of observing a sibship of size s containing r "affected" becomes $\binom{s-1}{r-1} p^{r-1}q^{s-r}/(1 - q^{s-1})$ which would yield the estimation equation as

$$\frac{R - N}{p} = \sum_s \frac{(s - 1)n_s}{1 - q^{s-1}}$$

These expressions are modifications of our previous formulas (8) and (9). Haldane (8) also has given variances of such estimates.

Another modification is the case in which a sibship is recorded in Table 1 only when it contains at least 1 "affected" *and* at least 1 "normal" member, so that both extreme classes of the binomial distribution are absent. The correction factor for these double truncated binomial series is obviously $1 - q^s - p^s$ (Finney (3)). The procedure for arriving at an estimation equation is exactly the same as before.

In summary, we may remark that though the problem of segregation of recessives originated from considerations of genetics, it is soon realized that it is really a general problem of estimating p and q of a binomial distribution when certain extreme classes are unknown.

II. THE SEVERITY OF AN ABNORMALITY

C. C. LI, *University of Pittsburgh*

IN THE FOREGOING chapter we have based our discussions on the assumption that individual traits can be diagnosed and classified into 2 clearcut groups, the "normal" and the "affected." But frequently more critical examination reveals great variance in degree of affectation. Some "normal" individuals may have a mild, almost undetectable manifestation of the trait, while some recognized "affected" individuals manifest the trait to a lesser degree than others. Also to be considered is the time or age of onset. It is important to make these distinctions in the collection of family records and in subsequent genetic analysis. Therefore, by classifying the traits of individuals into 3 groups (normal, semiaffected and affected) instead of 2, we will learn more about their genetic nature.

Before we delve into investigations on the severity of an abnormality, we must take into consideration the various possibilities that could produce the affectation.

First is the possibility that the trait is controlled by a single pair of genes. In this case we can distinguish heterozygotes (*Aa*) from the 2 corresponding types of homozygotes (*AA* and *aa*). When there is only 1 pair of genes to be considered it makes no real difference which allele is dominant over the other. Actually, neither can be considered truly dominant or recessive, as Table 5

TABLE 5

GENOTYPE	*AA*	*Aa*	*aa*
	Manifestations of Trait (Phenotype)		
Character dominant	severe	mild	normal
Character recessive	normal	mild	severe

shows. The terms "partial" and "semidominant" are sometimes applied, or *A* and *A′* are used to denote the alleles in order to demonstrate the nearly equal value of the genes. However, we may prefer for the sake of convenience to refer to a certain trait as being dominant or recessive. Arbitrarily, we may select the dominant trait according to: the phenotype of the heterozygote, or which trait was first discovered.

The second possibility is that the trait is caused by a set of multiple alleles with a certain scheme of dominance between them. This is very difficult to demonstrate when we are considering only 3 distinguishable phenotypes. (The established case of multiple alleles for human blood groups is based on 4 phenotypes.) There is really no well established case with respect to inheritance of human traits or diseases based on this multiple allelic hypothesis. The hypothesis, however, always remains a possibility.

The third possibility is that the trait is controlled by 2 or more pairs of genes. Owing to the multiplicity of genotypes and the various possible kinds of interactions between the different pairs of genes, even the 2-pair hypothesis is difficult to demonstrate. It can be done only when the pedigrees of families showing this trait are quite complete and comprehensive. As we have stated, if there were more than 3 distinguishable phenotypes, the demonstration would be much easier.

Now we are ready to cite several recent investigations on the subdivision of affectation according to degree or time of onset. We shall see how this division has led to new and more accurate genetic interpretation.

THE HETEROZYGOUS-HOMOZYGOUS HYPOTHESIS

Let us consider the *sickle cell trait,* referred to in medical literature as sicklemia or drepanocytosis. This is a condition in which the red blood cells, erythrocytes, assume the shape of a sickle or other bizarre forms when they are exposed to certain special conditions outside of the body. Approximately 7–8% of American Negroes show this trait. Soon after its discovery, it was determined that this trait is due to a single dominant autosomal gene. It seems to be harmless to the great majority of individuals who show it. However, a small proportion of individuals with this trait develop a severe, chronic, hemolytic type of anemia called sickle cell anemia. The mortality rate among patients with sickle cell anemia is much higher than that of persons with just the sickle trait. The usual explanation is that the dominant gene for sickling has a variable expression—more strongly expressed in some individuals than in others. Thus, the difference between the sickle cell trait and sickle cell anemia was at one time attributed to the irregularity in the degree of expression of the same gene, and the difference between exhibition of the trait and development of the anemia was not understood. It is true that in many family pedigrees there is indication that a gene may have various degrees of penetrance or variable expressivity, but this does not preclude other genetic interpretations.

Neel (20) suggested that the heterozygous condition results in sickle cell trait, whereas the homozygous condition results in sickle cell anemia. Neel's scheme follows: let *Sk* be the gene responsible for sickling the red blood cells

Genotype	*SkSk*	*Sksk*	*sksk*
Phenotype	anemia	sickling trait	normal

The patients with sickle cell anemia, then, receive the gene from both parents. (Note that if anemia is the trait we are considering, it would be equally correct to say that it is a recessive character because it manifests only when it is in the homozygous condition, whereas its corresponding heterozygous form is apparently normal, i.e., without the chronic anemia.) This newer hypothesis, then, states that the difference between sickle cell trait and sickle cell anemia is due not to the variable "expressivity" of the gene itself but rather to the different genotypes of the individuals concerned.

To prove the hypothesis that *SkSk* = anemia, we should find that both parents of an anemic patient must possess at least 1 *Sk* gene, and therefore both of them must show the sickle cell trait. Neel's chief evidence is that on testing 42 parents of 28 anemic patients every 1 of them was found to have the sickle cell trait. On the other hand, if anemia were merely a severe form of *Sksk*, then some of the parents of anemic patients would be of the genotype *sksk* and thus normal. The probability of finding 42 such parents, all having the sickle cell trait, would be very small indeed.

While it is true that examining the parents of anemic patients is the simplest and most practical procedure for proving the heterozygous-homozygous hypothesis, there are various other methods of approach which have yet to be made to confirm or to throw doubt on this theory. For instance, we may make a collection of families in which both parents have the trait (*Sksk* × *Sksk*) and determine the proportions of the various kinds of individuals among their offspring. Since the *Sksk* × *Sksk* type of mating can be identified by blood test without progeny-test, their total offspring should consist of $^1/_4$ normal + $^1/_2$ sickle + $^1/_4$ anemia. This can be calculated directly without resorting to the correction methods outlined in section I.

Since more than 92% of individuals in the population are normal, we know that the *Sksk* × *sksk* type of mating is more frequent than the *Sksk* × *Sksk* type. Therefore we should be able to obtain a fairly large collection of such *Sksk* × *sksk* families to determine if their offspring would consist of $^1/_2$ "normal" and $^1/_2$ with the sickle cell trait, but none with anemia.

A discussion of this trait is incomplete without at least a casual mention of the *Sk* gene frequencies in the general American Negro population. The proportion of the 3 genotypes (*SkSk*, *Sksk*, *sksk*) in a population in which mating is at random would be arrived at by applying the distribution of the Hardy-Weinberg law: let x be the gene frequency of *Sk* and y that of *sk* where $x + y = 1$.

$$x^2SkSk + 2xy\ Sksk + y^2\ sksk$$

Since we know that in the American Negro population there are approximately 8% of persons "affected," we may take $x = 0.04$ and $y = 0.96$ for the sake of illustration. Then in that population the proportion of the 3 genotypes would be

	ANEMIA		SICKLE		NORMAL
	0.0016		0.0768		0.9216
or	1	:	48	:	576

The frequency of the various types of matings mentioned above can immediately be obtained from these genotype proportions on the assumption that there is random mating in the population. Further, it will be seen that approximately 2% (1/49) of the individuals with sickle cells would be expected to have anemia.

There are many other human abnormalities which show varying degrees of "expressivity" in different individuals. It is true that the influence of some genes does not have uniform expression in all cases. It could vary depending on accidental environmental conditions, the genetic milieu of the individuals or some other unknown causes (this is demonstrated not only in human traits but in many plants and animals, as well), but, as we have seen, this does not mean that all variable phenotypes are always due to the irregular action of one and the same gene.

BIMODAL DISTRIBUTION OF VARIABLE PHENOTYPES

Segregation of a disorder into 2 clearcut groups (as in anemia vs. sickle cell trait) is not always possible. Many disorders show a continuous range of manifestations from very mild to very severe and/or with a time of onset from very early to very late in life. Moreover, the heterozygotes are conceivably more variable in expression than the homozygotes, and there might be an overlapping in expression between them. Even in such cases, however, a more detailed examination of the variable phenotypes might be fruitful.

For a long time genetic investigation has been directed toward a study of the prevalence of *asthma and hay fever* in the general

American population. Several hypotheses have been advanced in regard to the mechanism of heredity. The theory that this trait is due to a dominant autosomal gene fails to explain why in more than half of the pedigrees neither parent of "affected" patients is "affected," while the recessive theory fails to account for the fact that in some cases both parents are "affected" and yet some of their children are not.

It is, however, common knowledge that the age of onset of asthma and hay fever varies considerably and continuously from infancy to advance age. "Affected" patients cannot be divided into 2 distinct and separate groups, but the frequency distribution of the age of onset appears to show 2 modes (Fig. 1). The first

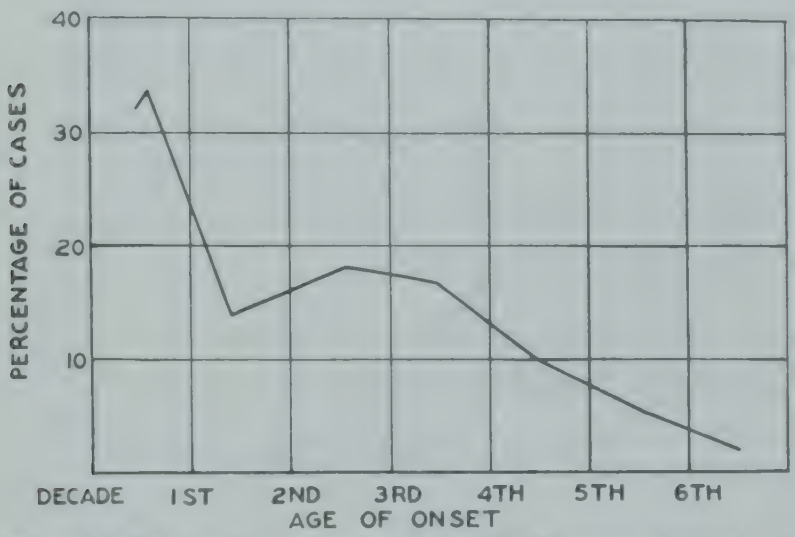


FIG. 1.—Bimodal distribution of asthma classified by age of onset. (Modified from Bray.)

mode is between 1 and 5 years of age; the second mode shows onset at about the third decade of life. This bimodal feature of the distribution strongly suggests that these so-called allergic patients constitute a heterogeneous group consisting of at least 2 subgroups: early-onset patients, and late-onset patients. The observed distribution (Fig. 1) would really be the sum of 2 separate distributions of the 2 groups of patients. There is some overlap between the first and second decade of life.

This and some other considerations led Wiener *et al.* (27) to suggest that these allergic patients are of 2 genotypes. Let *h* be the gene for developing the allergic condition and *H* be its normal allele. Wiener's hypothesis is then that the *hh* individuals will develop allergy "early" (onset before 10 years of age), whereas the *Hh* individuals, though the majority of them are "normal,"

will develop allergy "late," if at all. On the basis of this hypothesis, the form of distribution shown in Figure 1 would gain new significance. It will be noted that the right-hand portion of that curve (late-onset group) is much flatter, spreading over a wide range. We may assume from that that the *hh* individuals are less variable and will all develop the condition at about the same age, but the *Hh* individuals, when they do show allergy, are extremely variable with respect to the time of onset. This would agree with the general observation in genetics that heterozygotes usually vary more with environmental conditions than homozygotes.

What proportion of the heterozygotes will be "normal" and what proportion of them will develop these allergies? This question may be answered by a rough analysis of the allergic gene frequencies. Approximately 7% of the general American population manifest these allergic conditions. About one third of them (33–34%) belong to the early-onset group (see Fig. 1). Here we may apply the Hardy-Weinberg law. Let x be the frequency of this allergic gene h, and y that of its normal allele H, where $x + y = 1$. The proportion of *hh* individuals (the early-onset group) in the population is given by x^2, etc.

Genotype & Its Percentage			Phenotype & Its Percentage		
hh	x^2 =	2.33	Early onset	2.33	7%
Hh	$2xy$ =	25.86	Late onset	4.67	
			Apparently normal	21.19	93%
HH	y^2 =	71.81	Pure normal	71.81	
Total		100.00		100.00	

Now we see that 4.67/25.86 or 18% of the heterozygous individuals will develop these allergic conditions, but late. It is important to remark that this percentage is not constant but will vary from 1 family to another, depending on the environmental conditions. Among the 93% normal persons in the population, there are 21.19/93 or 22.8% heterozygous individuals who do not develop allergy although they are carriers of the allergic gene.

This hypothesis explains both cases: where 2 "normal" parents reproduce "affected" offspring, and where 2 allergic parents reproduce "normal" children. In the former case, even if the child develops allergy in early life (genotype *hh*), we see that his 2 parents may still be "normal." In the latter case, if the "normal" child is *Hh*, his 2 allergic parents may be $Hh \times Hh$ or $Hh \times hh$; even if the child is of the pure "normal" type (*HH*), his parents could still be allergic, both of the late-onset type. Wiener's explanation seeks to combine the heterozygous-homozygous hypothesis and the concept of variable expressivity.

Since it is assumed that these allergic patients consist of 2 different genotypes, it is necessary to analyze their parents separately in order to obtain further evidence supporting this hypothesis. It is not sufficient to know whether the parents of an early-onset allergic child are allergic or not; we have to know, in addition, whether they belong to the early- or the late-onset group. If the information is complete in this respect there are 2 courses we may take in order to test the hypothesis. We may determine the proportion of early- or late-onset type of allergic offspring among a certain type of parental combination and compare it with the expected proportion on this hypothesis; or, by a reverse process, we may calculate the frequencies of the various kinds of parents of a certain type of allergic patient and compare these frequencies with those actually observed.

In practice, however, various allowances have to be made in the calculations depending on the nature of the data at hand. For instance, in Wiener's own data almost all of the allergic patients are children and hence are presumably genotype *hh*. Among the mating of 2 apparently "normal" parents with at least 1 allergic child, the expected proportion of child patients is 25%. [Wiener *et al.* used a simpler formula than Haldane's in calculating the value of p by assigning q the value of .75, then solving for p. This simple procedure was originally conceived by Lenz (1929), and was extensively applied by Hogben (1931). It is, therefore, sometimes called the Lenz-Hogben method. It is really the first step of the iterative process in solving the equation (8). Historically, it was this method which led to Haldane's formula derived by the method of maximum likelihood.] However, about 10% of these children are under 4 years of age, so probably had not yet had time to develop the trait. The observed proportion, then, of allergic children would be below the theoretical 25%. If we assume that $^1/_4$ of this group of infants (the 10% group) would develop allergy later on, we would actually expect only $25 - {}^1/_4(10)$ or 22.5% allergic children. Wiener's observed proportion is 21.7%, which agrees closely with the expected.

The estimation of the proportion of late-onset allergic offspring presents even more difficulties because of the wide variation in age of onset. A person who is normal at the time of examination may develop allergy several years later. In making allowances for this, the distribution with regard to age of onset such as that shown in Figure 1 may be a valuable guide.

III. METHODS FOR ESTABLISHING THE GENETIC ROLE

C. C. LI, *University of Pittsburgh*

INVESTIGATION OF BLOOD RELATIVES and some of the features of the consequences of marriage between relatives provides the first method for establishing the genetic role.

BLOOD RELATIVES AND CONSANGUINEOUS MATING

Brothers and sisters are "bilineal" relatives, being connected through 2 independent persons, their mother and their father. This is to be distinguished from the "unilineal" relatives such as parent-offspring. Parent and child may be of the same genotype, but it is impossible for them to share the same 2 genes. For example, the probability that the child will receive the allele A of the A-a pair from his mother is $^1/_2$. Now, the child may receive the a gene from his father, thus becoming the same genotype as his mother. Bilineal relatives, however, may quite possibly share the same 2 alleles. Consider a child of the genotype AA', whose A is from his mother, while A' is from his father. The probability that his brother receives A from his mother is $^1/_2$. The probability that his brother receives A' from his father, as well, is $(^1/_2)^2 = {}^1/_4$. Thus, the probability that 2 sibs share exactly the same 2 alleles is $^1/_4$. This represents the simplest theory of unilineal and bilineal relationships. We will not go into further detail on that subject here. It is sufficient to remember that analysis of the relatives of propositi, particularly brothers and sisters, will often establish the hereditary fact of a trait, if not the mode of its inheritance.

To apply the Hardy-Weinberg law, suppose that x is the frequency of the rare recessive gene for a certain disease. The probability of finding an unrelated person showing the same disease as the propositus is x^2. This is usually a small number. But among the sibs of the propositus we expect $^1/_4$ of them to be affected by the same disease. So, in general, if the incidence of a disease is much higher among sibs or other near relatives than that among unrelated persons, we may presume that the disease may have a genetic component in its etiology.

The study of identical twins is most useful in establishing the fact that a trait is inherited, but it gives us no information whatever as to the mode of its inheritance. For instance, if a number

of identical twins showing a certain disease are studied, and it is noted that the severity and time of onset of the disease is much more similar than between sibs, the conclusion is warranted that not only the disease but also its severity and time of onset are genetically determined. On the basis of this evidence, one may say that 2 persons with the same disease but different times of onset are probably of 2 different genotypes. We do not, however, know whether the trait is dominant or recessive or due to multiple alleles or 2 pairs of genes, or sex-linked, etc.

Most matings in a population are between unrelated persons (random mating). Their total offspring would consist of x^2 homozygous recessives. When mating is between relatives, their total offspring consists of a higher proportion of homozygous individuals, depending on the nearness of their blood relationship. The closer the relationship, the higher the proportion of homozygous offspring. First-cousin marriages are by far the most important and common type of consanguineous matings in human populations.

Let F be the "coefficient of inbreeding" (28, 29) which measures the degree of consanguineous marriage. When first cousins marry, $F = {}^1/_{16}$. (We cannot go into detail here to explain the exact mathematical meaning of this index, but readers will find a comparatively simple account of the inbreeding problems in Chapter 14 of Li (15).) Now, if x is the frequency of the recessive allele, it may be shown that the proportion of homozygous recessives among consanguineous matings is increased from x^2 to

$$x^2 + Fx(1-x) = Fx + (1 - F)x^2$$

Thus, among the offspring of single first cousin marriages, there will be $\frac{x}{16} + \frac{15x^2}{16}$ recessives. If the value of x is small, as in the case of the rare abnormal gene, the proportion will be much higher than the original x^2 for random mating. Therefore, if we find a higher incidence of a certain disease among the offspring of consanguineous matings than in the general population, we may say that the trait probably has a recessive genetic basis. Conversely, we may investigate the parents of propositi. If there is a certain amount of intermarriage among the relatives of the parents of the patients, it is equally good evidence that the disease is probably due to a homozygous recessive condition.

With these few simple principles in mind, we may proceed to examine the results reported by Harris (9, 10) on the inheritance of *diabetes mellitus*. Like allergic asthma, this has long been a popular subject of inquiry by medical geneticists. The fact that this disease frequently occurs in several members of the same

family does not, in itself, warrant exact genetic interpretation. The study of identical twins has established the fact that hereditary factors are important in the etiology of diabetes but again sheds no light on its mode of inheritance.

One outstanding feature of diabetes mellitus is the very wide range of variation in respect to both severity and age of onset. It occurs much more frequently in later life than in early adult years. However, the early cases are usually severe and the late cases are usually mild. Umber (1939) and Lemser (1938), cited by Harris, observed that the severity of the disease was quite similar in both members of identical twins. This is the first evidence that the degree of severity of the disease is genetically determined. Therefore we may assume that the patients with early severe cases and those with late mild ones are of different genotypes.

Next, Harris (9) found a significant increase in the frequency of cousin marriages among the parents of the young diabetics but not among the parents of the diabetics of late onset. This not only suggests that the early and late cases are of different genotypes but indicates that early diabetes is probably due to the homozygous condition of certain genes.

In studying 3,827 sibs of 1,241 diabetic propositi of various age groups, Harris (10) found that 166 of the 3,827 or 4.34% were diabetic. Further, a significant correlation ($r = 0.695$) between the ages of onset was observed among the sibs who had developed the disease.

In interpreting this correlation, however, 1 allowance has to be made. Most of the unaffected sibs of young propositi were still quite young at the time of investigation and could later become diabetic, thus reducing the correlation of the time of onset mentioned earlier. To make this allowance, Harris calculated the number of sibs "expected" to be "affected." This was done in each of the various age groups of propositi, but disregarding a sib-sib correlation with respect to age of onset. His calculations are as follows:

If d is the number of diabetic sibs (among the 166 "affected") who developed the disease between the age t and $t + 10$ *years*, n is the number of sibs (among the 3,827 examined) who lived to the age $t + 5$ years and over (cumulative total up to that age); then the fraction d/n is the probability that a sib will develop the disease in the age interval t to $t + 10$ years, regardless of the age of onset of the propositi themselves. For each age group, the expected number of "affected" sibs is then obtained by multiplying the observed number of sibs (both "normal" and "affected") by the corresponding fraction d/n. Harris' results are summarized

in Table 6 which gives the pooled totals of a more detailed table.

It will be noted from Table 6 that the observed number of early-onset propositi with early-onset diabetic sibs is much higher than that expected on the assumption that they are independent of each other. Similarly, late-onset propositi tend to have late-onset sibs. This implies that the correlation between the age of onset of sibs, mentioned above, is real.

When all of the evidence was added, Harris was led to the plausible hypothesis that the early-onset severe type of diabetes mellitus is homozygous, whereas the late-onset mild type is heterozygous for a pair of autosomal genes. The heterozygotes vary a great deal in severity and time of onset, depending on environmental conditions and perhaps also on their genetic milieu.

TABLE 6*

Age at Onset of Propositus		Age at Onset of Diabetic Sibs			Total
		0–29	30–49	50–	
0–29	observed	26	15	1	42
	expected	9.63	5.42	0.88	15.93
30 and over	observed	15	45	64	124
	expected	31.37	54.58	64.12	150.07
Total		41	60	65	166

* Pooled totals from a more detailed table in Harris (10).

The homozygotes vary too, but to a lesser extent. The distribution of early and late cases probably overlaps more than in allergic disease and consequently is less clear-cut and more difficult to demonstrate by analysis of family records. The foregoing hypothesis covers all the facts mentioned: identical twins tend to have the same age of onset; young propositi tend to have young diabetic sibs; there are an excessive number of consanguineous marriages among the parents of young diabetics but not among the parents of patients with late onset. It is also consistent with the not infrequent occurrence of late diabetic patients among the parents (and other near relatives) of the young diabetics.

This is certainly a good working hypothesis, but to prove it conclusively, more critical data would be needed. The manifestation of this disease may be modified by genes of some other loci (modifiers); or the condition may be controlled by a multiple allelic series of genes. All these and other genetic factors deserve consideration.

However, in view of the above interpretations, diabetes mellitus, which had previously been considered as a recessive trait, we now

recognize as dominant. Indeed, this was the conclusion reached by Levit and Pessikova some years ago (1934, 1936; cited by Muller (19)). The phenomenon that the disease may sometimes skip a generation is due to the incomplete penetrance of the genes (usually below 20%). The same is true of allergic diseases described on page 22. Some of the potential diabetics may remain at such a low level of expression that they could not readily be detected. More refined means, such as blood sugar determinations, could be helpful in identifying them.

CONTROL GROUPS

We have seen several traits for which a simple mendelian mechanism can be postulated. There are those for which no hypothesis can be advanced except that the trait or disease is conditioned by the genetic constitution. (Owing to the wide range of variation in severity and age of onset, the inheritance of diabetes mellitus is on the borderline.) We will now examine some of these conditions.

To establish the genetic role in the etiology of a disease it is usually necessary to demonstrate that the incidence of the disease among the near relatives of the propositi is higher than that of the general population. This determination is accomplished by means of the "control" group, which is a group of individuals, selected at random, among people who are comparable to the propositi and their relatives in age, sex, occupation, economic conditions and other relevant respects.

The selection of an unbiased and really comparable control group is no easy task. Obviously, the control group has to be large for the incidence of the disease in it to be used as the measure by which the comparison is made. It is no exaggeration to say that the collection of data on an appropriate control group should be half of the total work in the investigation of the inheritance of a disease.

Furthermore, there are many "errors" involved in the 2 processes (establishing a control group and tracing the relatives of propositi) —errors that are made by both investigator and subject. These errors are usually automatic and unconscious, and because of this they are difficult to overcome. The diagnosis of the individuals in the control group tends to be less complete than among the relatives of patients: a person who knows he has a certain disease is more likely to be interested in searching out similar cases in his family. Either of these tendencies is capable of producing a positive error, and when they are combined the resulting figures may

lead to the false conclusion that the incidence of a disease among the relatives of patients is higher than that among the general population. Furthermore, we find that women usually give fuller information about their relatives than men. As a consequence, it would not be surprising to find a higher incidence of the disease among the relatives of female propositi than among those of males. All these possibilities of error show the necessity for taking elaborate precautions in the selection and examination of a control group.

There are various ways in which such a selection may be attempted. As mentioned before, a large number of comparable individuals chosen at random may be examined and the incidence of the disease among them determined, or, alternatively, we may use as control group, the corresponding relatives of an equal number of individuals comparable to the propositi. In studying the genetics of human cancer, for example, Waaler (1931, cited by Penrose *et al.* (21)) used the families of spouses of propositi as the control material. In order to prepare for the greater information given by women, we may select for each patient a control individual of the same age and sex whose near relatives are to be investigated. The considerations and precautions in selecting a satisfactory control group depend on the nature of the group of propositi as well as on the nature of the disease. It is the responsibility of the investigator to make a wise decision.

We shall cite an example to illustrate how to set up a standard of comparison with respect to the incidence of a disease. The fact that *peptic ulcer* is partially conditioned by hereditary factors has been demonstrated by the studies of identical twins (5, 22). Levin and Kucher (12) were the first to adopt a control group in the study of the genetics of the disease. They compared the near relatives of ulcer patients with 500 members of the general population and concluded that ulcer was 3 times as frequent among relatives of ulcer patients (this is cited by Doll and Buch (2)).

The study of Doll and Buch (2) largely followed the method of Levin and Kucher, but their control group was much larger. It consisted of 5,951 employed persons, of whom 4,871 were males and 1,080 females. These control individuals were very carefully examined. Diagnosis was established at a hospital by means of radiographic or gastroscopic examination or by operation, or there had been a history of frank hematemesis. Since we are primarily interested in the method of using a control group, rather than in the findings of the investigation itself, we will cite only the results concerning the 4,871 male controls (see left side of Table 7). Exactly the same thing was done with the 1,080 female

controls except that certain age groups were pooled because of the small number of individuals in each age group.

Doll and Buch also investigated the families of 300 ulcer propositi, using the same criteria for proof of an ulcer as those used in the control group. As the controls were living persons, the group for comparison was also confined to living relatives. It is possible that in a few cases ulcers were falsely diagnosed, but, at the same time, some were undoubtedly omitted because of ignorance on the part of the subjects interviewed. Therefore, the recorded number of ulcers among the relatives of propositi is, if anything, too low. Since we have cited only the male control group, the

TABLE 7.—Incidence of Ulcer (2)

Age	Male Controls			Brothers of Propositi		
	No. Seen	Proved Ulcers		No. Living	Proved Ulcers	
		No.	%		Observed	Expected
14-	199	0	0.00	6	0	0.00
20-	300	4	1.33	10	0	0.13
25-	1,128	28	2.48	63	2	1.56
35-	1,375	63	4.58	143	12	6.55
45-	1,089	80	7.35	153	31	11.24
55-	625	39	6.24	84	11	5.24
65-	155	9	5.81	62	4	3.60
Total	4,871	223	—	521	60	28.32

results obtained on the brothers of the 300 propositi are given for comparison (right side of Table 7).

Sex and age are the 2 most important factors affecting the development of an ulcer. Therefore, the control group is subdivided into males and females and according to age. For each group, the percentage of ulcer patients was calculated (column 4 of Table 7). The percentages for the control group are regarded as the "standard rates" of ulcers among the people of that age group, regardless of distribution of the disease among their families. The "expected" number of ulcer cases shown in the last column of Table 7 is obtained by multiplying the number of living brothers by the corresponding age standard rates. For instance, there are 143 brothers of propositi in the age group 35–44 whose ulcer standard rate is 4.58%. Therefore, we expect that among 143 brothers of that age there should be $143 \times 4.58\% = 6.55$ ulcer patients. Instead, the number of ulcer cases actually observed is 12. From the bottom line of Table 7 it is clear that the total number of "affected" brothers of propositi is much higher than the expected number based on the rates of the male control group. Indeed, it is $60/28.3 =$

2.1 times as many. The same phenomenon was observed when the "affected" sisters of propositi were compared with the female control group. The same was found true with respect to the fathers of the propositi.

In further analyzing their data, Doll and Buch observed that familial tendencies were more prominent in duodenal ulcer and in patients with early onset, but the differences were not statistically significant. There was a strong tendency for sibs to have ulcers in the same site, whether duodenal or gastric. These findings led to the conclusion that hereditary factors are important in determining the development of peptic ulcer. But, on the other hand, at the present stage of knowledge, no simple mendelian mechanism seems plausible.

GENERAL POPULATION AS A CONTROL

Some of the difficulties in obtaining a satisfactory control group were mentioned previously. To avoid these difficulties, it is possible to use the general population as the control group when relevant statistics are available. For studies in human *cancer*, for example, it proved to be convenient.

Many investigators were unable to demonstrate that the incidence among the relatives of cancer propositi is much higher than that in the general population; and the conclusion was reached that there is no hereditary transmission of the disease. In almost all these investigations, however, some uncertainty remained as to whether the incidence of cancer, ascertained in the control group, was correct. This led Penrose *et al.* (21) to use the published statistics of the general population as the control group in their study of the inheritance of human mammary cancer, thus dispensing with the almost insuperable task of collecting a large and appropriate control group.

The method of using the general statistics as control is similar to that described in the preceding article. Penrose *et al.* tabulated (from the population of England and Wales) death rates for both females and males in every age group through the years 1911–45. The results were used as the "standard rate" from which was calculated the "expected" number of deaths due to mammary cancer among the relatives of propositi. For each age group (e.g., 50–55 years old) and each year group (e.g., 1920–24) there is a standard death rate for females (or males). The deaths of the patients' relatives (mothers, sisters, fathers, brothers, daughters and sons) were recorded, again by age group and year group. Then the total observed number of deaths due to this disease was

compared with the total expected based on the standard rates. The various death rates for other types of cancer were also determined. Penrose's results concerning the mothers and sisters of 510 female propositi are summarized in Table 8.

It is clear that for both mothers and sisters of the propositi there is a highly significant excess in observed deaths from mam-

TABLE 8

Total Deaths	Deaths Due to Mammary Cancer		Deaths Due to Other Cancers	
	Obs.	Exp.	Obs.	Exp.
Mothers 406	25	11.12	51	49.23
Sisters 307	23	6.97	19	25.23

mary cancer over that expected, but there is no significant divergence from expectation with respect to deaths from other types of cancer. A similar calculation for the fathers and brothers of propositi showed no observed excess of deaths from the disease over that expected. As to the daughters and sons of propositi, they are too few to give reliable results. It seems plausible, therefore, that a *specific* genetic agent responsible for the disease might be inherited through the maternal line. As to the question of transmission by way of maternal milk (as is the case in mice), the data are incomplete and the results inconclusive. Penrose *et al.* also observed that there is strong similarity among "affected" sisters with reference to the age of onset and the site (right or left) of the initial lesions. It is needless to add that the genetics of human cancer will remain a challenging problem for medical geneticists for some time to come.

IV. LINKAGE VERSUS ASSOCIATION

C. C. LI, *University of Pittsburgh*

As stated earlier, our discussion on genetic linkage will be brief. For a more complete but concise account of the methods of measuring linkage, one may refer to Chapter X of Mather's monograph (17).

The term "linkage" in genetics is a technical one. It means that 2 loci (the site of genes) are located on the same chromosome. Thus the 2 genes are linked by the common chromosome on which they locate, to be separated only by a process of breaking-and-rejoining between the chromatids known as "crossing-over."

For example, any allele of the *A*-*a* series may be linked with any member of the *B*-*b* series. When there are only 2 alleles in each series, there are 4 possible kinds of chromosomes for the 2 pairs of genes, namely, *AB*, *Ab*, *aB*, *ab*. When there are 3 alleles to the *A* series and 2 alleles to the *B* series, there will be 6 possible kinds of chromosomes with respect to the 2 loci, and so on.

When the 2 pairs of genes are located on different pairs of chromosomes, we say that they are independent, and we refer to the double heterozygous individuals as *AaBb*. But when they are linked, the reference is *AB/ab* or *Ab/aB*, depending, of course, on which 2 genes are situated on the same chromosome in this particular zygote. Obviously, when 1 of the pairs is homozygous, it is irrelevant to make this distinction, and the form *Aabb* is just as good as *Ab/ab*. In either case, this zygote will produce half *Ab* and half *ab* gametes. This, however, is not the case with double heterozygotes. The *AB/ab* zygote will produce proportionally more *AB* and *ab* gametes than the cross-over type of *Ab* and *aB* gametes. Similarly, the *Ab/aB* individuals will produce more *Ab* and *aB* gametes than the cross-overs, *AB* and *ab*.

To demonstrate, let x, y be the frequencies of the genes A, a and let u, v be the frequencies of the genes B, b in the general population where $x + y = 1$ and $u + v = 1$. In a random mating population which is in condition of equilibrium, the proportions of the 9 genotypes with respect to the 2 pairs of factors, *whether linked or independent*, are entirely determined by the values of gene frequencies and are given by the following symbolic product

$$\begin{pmatrix} AA & Aa & aa \\ x^2 + 2xy + y^2 \end{pmatrix} \begin{pmatrix} BB & Bb & bb \\ u^2 + 2uv + v^2 \end{pmatrix} \qquad (10)$$

This can readily be recognized as a simple extension of the Hardy-Weinberg law with which we are now familiar. The proof of this important theorem is unfortunately too long to be given here, but readers will find a comparatively simple account of this property in Chapter 8 of Li (15). When the expression (10) is multiplied out and the 9 terms are arranged in the form of a 3×3 table, it will be clear that there is no genotypic correlation among them, even though they are "linked." (Of course, this is evident from the product form itself.)

If A is dominant over a, and B is dominant over b so that there are only 4 distinguishable phenotypes with respect to 2 traits, the phenotypic proportions in the general population can readily be obtained by collecting and summing the terms showing the same phenotype in the expansion of (10). It will be found that the 4 phenotypic proportions in equilibrium are as shown in Table 9. These proportions hold true whether the 2 pairs of genes

TABLE 9

	B–	*bb*	Total
A–	$(1 - y^2)(1 - v^2)$	$(1 - y^2)v^2$	$1 - y^2$
aa	$y^2(1 - v^2)$	y^2v^2	y^2
Total	$(1 - v^2)$	v^2	1

are independent or linked. It is evident from Table 9 that there is also no phenotypic correlation between the 2 characters. The proportion of each phenotype is given by the product of the corresponding marginal totals, and they in turn are the proportions of the phenotypes with respect to 1 character. So we reach the important conclusion that even if the 2 pairs of genes are linked, there will be neither genotypic nor phenotypic correlation in the general population. Genetic linkage, if and when it exists, cannot be detected or demonstrated by ordinary correlation studies between 2 characters.

Expanding (10), we find that the proportion of double heterozygotes (*AaBb*) in the general population is $4xyuv$. In an equilibrium population, the double heterozygotes of the "coupling" type, *AB/ab*, are as numerous as those of the "repulsion" type, *Ab/aB*, the proportion of each being $2xyuv$. Since the 2 types of double heterozyotes are equally numerous in the general population, it follows that the 4 kinds of gametes produced by these double heterozygotes will also be equally numerous *whatever the*

strength of linkage between the 2 genes. The scheme in Table 10 will make it clear. So we see that the total gametic output of the double heterozygotes is $^1/_4$ of each of the 4 kinds of gametes.

TABLE 10

Type of Zygote	Gametes Produced			
	AB	*Ab*	*aB*	*ab*
Coupling: *AB/ab*	*n*	1	1	*n*
Repulsion: *Ab/aB*	1	*n*	*n*	1
Total	$1 + n$	$1 + n$	$1 + n$	$1 + n$

This is precisely the case with independent genes where the double heterozygotes *AaBb* produce an equal number of each kind of gamete.

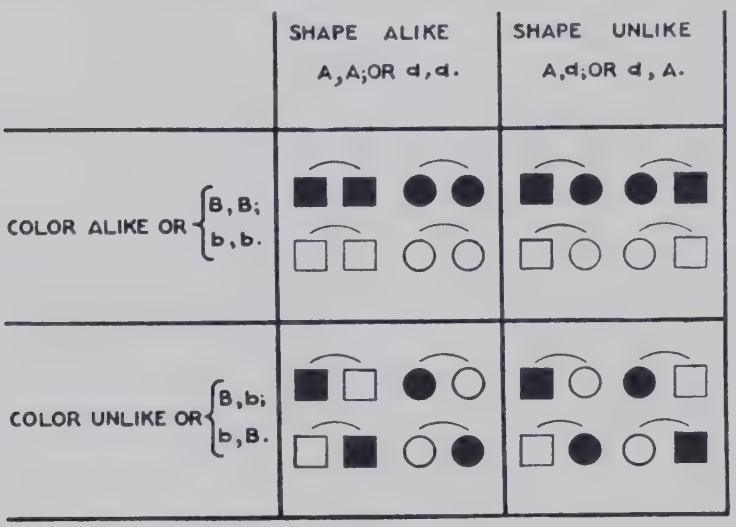
The foregoing discussion not only makes it clear that genetic linkage cannot be detected from the general population but indicates where we should look for the effects of linkage. Evidently, the linkage effects will be shown in the sibships with at least 1 double heterozygous parent. The simplest type of family which would enable us to detect and measure linkage is *AaBb* × *aabb*, which is equivalent to "back-cross" in plants and animals. Among the offspring of this type of parental combination, either the *AB* and *ab* children are more numerous than *Ab* and *aB* or vice versa, according to whether the double heterozygous parent is of the coupling or repulsion type. (It will be recalled that he may be of 1 type or the other.) If he is of the repulsion type, then there will be more children of the types *Aabb* and *aaBb* than *AaBb* and *aabb*. That is to say, the majority of the offspring will possess 1 character but lack the other (hence the term repulsion). In other words, there will be *negative* correlation between the 2 characters within such sibships. On the other hand, if the double heterozygous parent is of the coupling type, the reverse will be true, and positive correlation will be observed within the sibships. These 2 contrasting kinds of sibships are equally numerous in the general population.

The next important types of families suitable for linkage studies are *AaBb* × *Aabb* (or *aaBb*) and *AaBb* × *AaBb* (the latter being equivalent to raising F_2 by selfing or *inter se* crossing of F_1 individuals in plants and animals). The same general phenomena also hold true for these types of mating.

The genetic effects of linkage can also be studied by comparing the members within sibships alone, without knowing the types of their parents. This is based on the fact that when linkage exists

between 2 loci, not only the number of sib-pairs which are alike with respect to both of the 2 characters (positive correlation) will be more than it would be if the genes were independent, but the number of sib-pairs which are unlike with respect to both characters (negative correlation) will also be more numerous than the independent case (Fig. 2).

We have presented this brief discussion of linkage to emphasize that correlation between 2 characters in the general population is



SIB-PAIRS FOR ESTIMATING LINKAGE
1ST CHARACTER (A, a) SHAPE: A□ ,a○
2ND CHARACTER (B,b) COLOR: B-BLACK, b-WHITE

FIG. 2.

never an indication of genetic linkage and to point out where the genetic effects of linkage can be found.

The association between 2 characters in the general population may be due to a variety of causes. It may be that the same gene or set of genes affects both characters, or that 1 character is a natural physiologic consequence of the other. It may be that the population has been subdivided into racial groups so that it is no longer a homogeneous random mating community; differential inbreeding also increases the degree of association between traits in the general population. Finally, the association may be due simply to certain bias in the process of collecting the sample.

It might be added that the difficulties of studying linkage in human populations are due not only to the equilibrium properties we have described but also to the small number of offspring in each family. Certain genetic combinations simply have no chance of being identified and segregated. If each sibship consisted of 4

or more members, the task of measuring genetic linkage would be much easier and more reliable.

Comment by Howard Levene

This section should have great value in introducing the medical man to recent developments in the study of human genetics and should thus help increase the number of useful studies of the many morbid conditions having a genetic background. The discussion of the theoretical background of these methods is clear and understandable, and the nonmathematician may skip much of the more forbidding mathematics without losing the thread of the argument. On the other hand, the mathematics is there for those who wish to follow it.

In part I the extension of Weinberg's sib-method to varying family sizes

$$p = \frac{\sum_s a_{rs} r_s (r_s - 1)}{\sum_s a_{rs} r_s (s - 1)} \tag{a}$$

might have been given. This estimate is very easy to calculate but is less efficient than the use of equation (8). The estimate from equation (a) may be used as the first guess in the trial and error solution of equation (8), where successive guessed values of p are substituted into both sides until the equality holds to the desired accuracy. (Any mathematical friend of an investigator can explain Newton's method of solution, which is less laborious.)

I am particularly glad to see part IV, which should go far to clear up the rather common confusion between association of traits and linkage of genes.

The author has wisely restricted himself to a few important problems, rather than try to cover the whole field in a short space. Within his chosen limitations he has done an excellent job of presenting the modern developments.

REFERENCES

1. Dahlberg, G.: *Mathematical Methods for Population Genetics* (New York: Interscience Publishers, Inc., 1948).
2. Doll, R., and Buch, J.: Hereditary factors in peptic ulcer, Ann. Eugenics 15: 135–46, 1950.
3. Finney, D. J.: The truncated binomial distribution, Ann. Eugenics 14: 319–328, 1949.
4. Fisher, R. A.: The effect of methods of ascertainment upon the estimation of frequencies, Ann. Eugenics 6: 13–25, 1934.
5. Freeman, A. G.: Peptic ulceration in identical twins, Brit. M J. 1: 765, 1947.
6. Haldane, J. B. S.: A method for investigating recessive characters in man, J. Genetics 25: 251–255 1932.

7. Haldane, J. B. S.: The estimation of the frequencies of recessive condition in man, Ann. Eugenics 8: 255–262, 1938.
8. Haldane, J. B. S.: A test for homogeneity of records of familial abnormalities, Ann. Eugenics 14: 339–341, 1949.
9. Harris, H.: The incidence of parental consanguinity in diabetes mellitus, Ann. Eugenics 14: 293–300, 1949.
10. Harris, H.: The familial distribution of diabetes mellitus: A study of the relatives of 1,241 diabetic propositi, Ann. Eugenics 15: 95–119, 1950.
11. Hogben, L.: *An Introduction to Mathematical Genetics* (New York: W. W. Norton, 1946).
12. Levin, A. E., and Kucher, B. A.: On the clinical-genetical differentiation of ulcerous diseases, Proc. Maxim Gorky Med.-Genet. Res. Inst. (Moscow) 4: 181, 1936 (in Russian, original not consulted).
13. Levit, S. G., and Pessikova, L. N.: The genetics of diabetes mellitus, Proc. Maxim Gorky Med. Biol. Res. Inst. (Moscow) 3: 132–147, 1934 (Russian with English summary).
14. Levit, S. G., and Pessikova, L. N.: Is diabetes mellitus caused by a good dominant gene? Proc. Maxim Gorky Med.-Genet. Res. Inst. (Moscow) 4: 149–158, 1936 (Russian with English summary).
15. Li, C. C.: *An Introduction to Population Genetics* (Peiping, China: National Peking University Press; and Corvallis: Oregon State College Cooperative Association, 1948).
16. Mandeville, L. C.: Congenital absence of permanent maxillary lateral incisor teeth: A preliminary investigation, Ann. Eugenics 15: 1–10, 1949.
17. Mather, K.: *The Measurement of Linkage in Heredity* (New York: John Wiley & Sons, Inc., 1951).
18. Muller, H. J.: Progress and prospects in human genetics, Am. J. Human Genetics 1: 1–18, 1949.
19. Muller, H. J.: Our load of mutations, Am. J. Human Genetics 2: 111–176, 1950.
20. Neel, J. V.: The inheritance of sickle cell anemia, Science 110: 64–66, 1949.
21. Penrose, L. S.; Mackenzie, H. J., and Karn, M. N.: A genetical study of human mammary cancer, Ann. Eugenics 14: 234–266, 1949.
22. Riecker, H. H.: Peptic ulcer in identical twins, Ann. Int. Med. 24: 878, 1946.
23. Snyder, L. H.: Old and New Pathways in Human Genetics, in *Genetics in the 20th Century* (New York: Macmillan Company, 1951).
24. Stern, C.: *Principles of Human Genetics* (San Francisco: W. H. Freeman, 1950).
25. Strandskov, H. H.: Genetics and the origin and evolution of Man, Cold Spring Harbor Symposia 15: 1–11, 1950.
26. Weinberg, W.: Mathematische Grundlagen der Probandenmethode, Ztschr. Abstgs. und Vererbgsl. 48: 179–228, 1927.
27. Wiener, A. S.; Zieve, I., and Fries, J. H.: The inheritance of allergic disease, Ann. Eugenics 7: 141–162, 1936.
28. Wright, S.: Coefficients of inbreeding and relationship, Am. Nat. 56: 330–338, 1922.
29. Wright, S.: The Effects of Inbreeding and Crossbreeding on Guinea Pigs: III. Crosses between Highly Inbred Lines, U.S. Department of Agriculture Bulletin no. 1121, 1922.

SECTION II

Methods in Environmental Medical Research

ASSOCIATE EDITOR—*Ray G. Daggs*

INTRODUCTION

THE STUDY OF the effects of various natural and man-made environmental situations on the human or animal organism involves many of the same methods and principles used in other medical science studies. However, certain factors must be given special consideration in environmental studies. In most instances the organism is studied as an integral unit rather than as isolated systems or tissues. One of the exacting requirements is to know the conditions of the environment so as to assess properly the amplitude of the stressful stimulus. For this reason it was deemed advisable to include in this section a somewhat detailed discussion of the methods of measuring climatic variables.

There are 2 general situations under which environmental studies are usually made laboratory and field conditions. There has been no specific attempt to differentiate clearly the methods for each situation since the approach often depends on the investigator's ingenuity and the particular facilities available. Environmental studies may include any or all of a whole gamut of determinations such as energy exchanges, water and electrolyte balances, cardiovascular responses, endocrine responses, nutrition and intermediary metabolism, neuromuscular functions, sensory responses, pain, fatigue, psychologic alterations and so on.

Many of the standard methods that may be used in environmental studies have been covered in previous volumes of this series and thus have not been repeated here. Blood flow measurements have been discussed in Volume 1; measurement of respiratory and blood gases in Volume 2; physical work and strength tests in

Volume 3, and fluid and electrolyte exchange in Volumes 4 and 5.

Methods for the study of the effects of high and low ambient pressures have not been included here nor methods for studying many unusual environments such as ionizing radiation, etc. Attention has been given to the applicability of methods to physiological research under the more usual naturally occurring climatic conditions. Since space is limited, consideration was given to those methods and procedures that are most common to all studies of the effects of climatic conditions, and those that have been covered adequately in other volumes of this series were omitted.

—Ray G. Daggs.

MEASUREMENT OF CLIMATIC VARIABLES

RICHARD L. PRATT *and* AUSTIN HENSCHEL, *Quartermaster Climatic Research Laboratory, Lawrence, Mass.*

THE DESIGN OF any experimentation involving human subjects must include provisions for recording the weather conditions in contact with the subject (microclimate). Such precautions are necessary because the physiologic, biochemical and psychologic responses of man are greatly influenced by air temperature, wind speed, humidity and radiation. Actually, under extreme conditions the responses of man may be governed by the ambient air conditions and, consequently, may entirely mask or override responses to other factors under consideration. Unless proper control of, or corrections for, the environmental stimuli are made the data derived from the experimentation become difficult to interpret and impossible to compare with results obtained under a different set of conditions. Unfortunately, only too frequently the results of otherwise well conceived and executed studies have proved to be of only limited usefulness because the climatologic conditions at the time of the testing were not recorded.

I. Air Temperature

The temperature of the air surrounding a man is frequently the most important measurement to be made when the environmental factors affecting man are being considered. Depending on the precision required by the experiment, there are various methods for measuring the environmental temperature. In all cases, certain precautions are necessary to insure results that are reasonably accurate and not misleading. The exposure of the temperature-sensitive element, the effect of lag, the degree of sensitivity and accuracy of the instrument and the position of the instrument relative to the man must all be considered.

1. *Exposure of the instrument.*—If a temperature-sensitive device such as an ordinary thermometer is placed in the air, it will eventually reach the same temperature as the air around it only if there is no gain or loss of heat by radiant energy exchange with some object which is warmer or cooler than the air. For example, a thermometer placed in sunlight will absorb part of the short wave radiation falling on it. Heat will be supplied to the thermometer and the temperature it registers will be higher than that of the sur-

rounding air. If the thermometer is well ventilated, the heat may be removed by conduction almost as fast as it is received. Under these conditions of high ventilation, there will be less difference between the true air temperature and the temperature of the thermometer, which is, of course, the temperature that is read provided the thermometer is accurately calibrated.

Radiant exchange may also occur whenever the thermometer is exposed to an object or the sky which is at a temperature different from the air around the thermometer. This generally creates less error than exposing it to direct sunlight, but the error may be of considerable magnitude. For this reason, thermometers and other temperature-sensitive instruments are usually shielded and ventilated by one means or another to reduce the effects of radiant exchange. A shielded instrument must be given sufficient ventilation; otherwise little or no benefit will be derived. For meteorologic and climatologic purposes an "instrument shelter" is used, not only to protect the instruments from radiation but also to protect them from rain and snow.

INSTRUMENT SHELTERS.—The standard instrument shelter is a louvered wooden box with a double roof that is painted white to reflect solar radiation. It is mounted on wooden legs so that the box is approximately 5 ft above the ground. For any extended field study, it is usually worth while to use such a standard instrument shelter for part of the temperature measurements if for no other reason than to make measurements comparable with climatologic records made all over the world. Usually, however, it will be necessary to make other measurements in addition to those obtained from instruments in the instrument shelter. Small shields, usually constructed of aluminum, are often used with thermometers and other measuring devices that are not housed in an instrument shelter.

As already mentioned, increased ventilation reduces the difference between the reading of the measuring device and the true air temperature. Reduction of the size of the measuring device acts in the same manner as increase of the ventilation. For example, a very fine thermocouple requires very little ventilation even in direct sunlight for reasonably accurate results.

2. *Effect of lag.*—Generally speaking, the larger (one with greater specific heat) the temperature-sensitive element, the greater the lag of the instrument. In other words, a large element requires more time to come into equilibrium with the air than does a small one. When the air temperature is fluctuating, an instrument with considerable lag will partially average the small changes in temperature but will never reach complete equilibrium. This in no way

ffects the accuracy of the instrument, but 2 accurate instruments vith different lag characteristics will not agree during periods of luctuating or changing temperature conditions.

3. *Position of temperature-sensitive device.*—In measuring air emperature for physiologic purposes, the standard instrument helter is not always sufficient because the temperature at only 1 evel is obtained. Frequently there are large differences in air emperature between the ground level and the region 4–6 ft above ground, which comprises the main environment of a standing man. For example, during July, 1952, in the desert near Yuma, Ariz., he temperature 6 in. above the ground was frequently 5–10° varmer than at the height of the instrument shelter, while the tem- perature at the surface of the ground averaged 35° warmer during he early part of the afternoon. This variation is even greater on calm sunny days (there was an average wind of 7 mph during the period mentioned) and is not confined to desert areas. For an ex- treme example, at 1008 hours on June 27, 1948, on the edge of a swamp near Ladd Air Force Base in Alaska, the temperature ½ in. above the ground was 126 F, at 1 ft 93 F, and at 8 ft above he ground 86 F; wind speed at this time was less than 1 mph. On the other hand, on cloudy, windy days little or no difference in temperature will be observed between the gound level and 5 or 6 ft above the ground. On clear, calm nights there is often a strong in- version of temperature (colder near the ground) so that the tem- perature of the air surrounding a man sleeping on the ground night be many degrees less than that indicated by a recording in- strument at 5 or 6 feet. This is also often true indoors in heated laboratories and observation rooms.

It is apparent that the environment, as far as air temperature is concerned, may be quite different at different levels and in certain circumstances. It is quite important that the temperature be meas- ured not at just any level but at the level occupied by the man.

In addition to level above the ground, there is often considerable difference in temperature from place to place. This difference is usually most noticeable where the terrain is variable. Small de- pressions are likely to be much cooler at night than nearby higher ground. Large-scale inversions may increase the temperature in a relatively short distance as one goes up a hill. There also may be large-scale temperature gradients near large bodies of water. In consequence, it is advisable to make the temperature measurements as close to the subjects as is practical.

4. *Accuracy and sensitivity.*—Air temperature often fluctuates as much as 2 or 3° in a period of a minute at 1 point; it varies greatly with altitude, particularly in the first 5 ft above the ground, and it

varies laterally particularly due to differences in terrain. Since air temperature varies with time, elevation and space, it is impractical to attempt to determine the environmental temperature of a man during a physiologic experiment in the field to less than 1° F, and probably even this accuracy is unobtainable under many conditions. Most instruments on the market for measuring temperature are therefore sufficiently accurate for the purpose, provided they are properly exposed. Some may be more sensitive than others, some may be more accurate, but, generally speaking, no better results will be obtained by using a precision thermometer accurate to 0.01° than by using 1 accurate to only 1°. To report ambient temperatures in the field to less than 1° F is merely giving a false impression of accuracy.

INSTRUMENTS FOR MEASURING TEMPERATURE

1. LIQUID-FILLED THERMOMETERS, INDICATING TYPE

These include the ordinary mercury-filled or alcohol-filled indicating thermometers. The mercury-filled thermometers are usually more accurate and maintain their accuracy better than those filled with alcohol but cannot be used at temperatures below −38° F, for at that temperature the mercury freezes. When an alcohol-filled thermometer is used, care must be taken to see that none of the fluid has evaporated from the surface of the liquid and condensed somewhere near the top of the thermometer. If this occurs (as it often does) and goes unnoticed, the top of the main column of liquid will be too low, possibly giving grossly inaccurate readings. Sometimes if jarred the column will separate, forming 1 or more bubbles in the main column, causing the top of the column to be too high. If either occurs, the column may usually be rejoined by tapping the bulb end on something soft enough not to break the bulb but hard enough to jar the thermometer. Slow, repeated tapping is best and may take 5 or 10 min to rejoin the column. If after 10 min no apparent improvement is noted, and the glass has not yet been broken, place the bulb in crushed dry ice or salt and ice mixture and keep the stem warm. The alcohol will evaporate from the stem and condense in the bulb. This will work, but it may be slow. This type of temperature-measuring device must be either shielded from radiation or highly ventilated or both. High ventilation is often accomplished by swinging the thermometer rapidly, as is done with the sling psychrometer. Merely holding such a thermometer in the hand, even if shaded by the body, is not sufficient (errors of 3 or 4° F are common).

2. *BI-METALLIC THERMOMETERS AND THERMOGRAPHS*

The temperature-sensitive element of a bi-metallic thermometer is a compound strip of metal made of 2 different metals welded together. The metals used have different coefficients of expansion so that with a change in temperature the strip is deformed. This deformation may be connected in such a way that a pointer will indicate the temperature or it may be the sensitive element of a thermograph. Like liquid-filled glass thermometers, proper exposure must be given so that the effects of radiation are not important. Well-made bi-metallic thermometers may be accurate, but many of the cheap indicating thermometers of this type may be far from accurate. When properly made and calibrated, a bi-metallic strip is often used to actuate a thermograph with good results.

Thermographs may also be activated by a Bourdon Tube, which is a bent metal container of elliptical cross-section filled with a liquid. The liquid expands when heated; the increased pressure inside the container tends to straighten the element which activates the pen arm. This type of thermograph is equally accurate and is a very convenient instrument to use. It should be shielded not only from radiation but also from rain and snow, so that a standard instrument shelter is almost a necessity for field use. Even in an instrument shelter, blowing snow often covers the bi-metallic strip or Bourdon Tube, causing erroneous readings, particularly when the ambient air is above freezing. This, of course, applies equally to all thermometers.

Remote recording thermographs are sometimes made with a sensitive element connected by capillary steel tubing to the recording mechanism. In this type, a bulb (the sensitive element) is filled with liquid, gas or vapor which expands when heated. The increased pressure is transmitted to the recording mechanism by means of capillary tubing. The pressure change is calibrated on a temperature scale. Considerable accuracy can be achieved, but such instruments are generally used at fixed installations. Because the bulb is usually quite large, it must be shielded from radiation. There is considerable lag due to the size, which renders the instrument quite slow in its reaction.

3. *MAXIMUM AND MINIMUM THERMOMETERS*

These are often useful for obtaining a rough average of temperature over a 24 hr period by taking the average of their readings. This method is often used in climatologic work and usually agrees rather well with the average of hourly observations. However, on any particular day, there may be considerable difference between

the average of the hourly temperature observations and the average of the maximum and minimum temperatures. For short tests they are of little value except to record the extremes of temperature. They are sometimes used to make corrections on a thermograph chart, but care must be taken owing to the different lag characteristics between the 2 types of sensitive elements. The maximum thermometer is likely to read higher than the highest point on the thermograph due to the greater lag of the thermograph. The minimum thermometer may read lower, but this is less likely for the time of the minimum temperature is usually during a period of stability when rapid fluctuations in temperature are unlikely.

The commonest *maximum thermometer* is a mercury-filled glass tube type thermometer with a fine constriction in the bore near the bulb. The remaining space in the tube is a vacuum. When the mercury in the bulb expands, it is forced past the constriction; when it contracts, the mercury which has been forced past the constriction remains there so that the maximum temperature may be read. Such thermometers are re-set by swinging them with the bulb down in such a manner that centrifugal force forces the mercury past the constriction into the bulb. A clinical thermometer is just a maximum thermometer with a limited range.

Minimum thermometers are usually of the liquid in glass type, with alcohol used as the liquid for transparency. Inside the column of liquid in the bore of the thermometer is a black glass dumbbell-shaped rod. As the temperature falls, surface tension at the top of the alcohol column pulls the dumbbell down. When the column rises, the alcohol flows past the dumbbell, leaving it at the point of minimum temperature. To re-set, the thermometer is tipped bulb end up until the dumbbell returns to the "top" of the column of alcohol. Both maximum and minimum thermometers are exposed in an almost horizontal position—the maximum with bulb slightly higher than the top, and the minimum horizontal. Suitable protection against radiation must be employed.

Another type of *maximum-minimum thermometer* is made which indicates both maximum and minimum temperatures on the same instrument. Glycerin is used as the liquid which expands or contracts with changes in temperature. The glycerin bulb is attached to 1 end of a U-tube and an expansion chamber is at the other end. The U-tube is partially filled with mercury which moves as the glycerin expands or contracts. Small steel indicators float on the mercury on both sides of the U-tube. As the mercury is displaced in either direction the indicators are raised. They remain at their highest point by means of small springs until re-set by drawing them back to the level of the mercury with a magnet. One side is cali-

rated to read minimum temperature and the other the maximum mperature.

THERMOCOUPLES

When 2 different metal wires, such as copper and constantan, re joined at both ends a current will flow if the 2 junctions between ne metals are at different temperatures. When 1 of the junctions eference junction) or thermocouples is at a known temperature, ne temperature of the other may be determined by measuring the ectromotive force produced. The reference junction is usually aintained at 0° C by keeping it immersed in an ice and water ath. With copper-constantan thermocouples about 40 μv are oduced per degree differential in temperature between the refer- nce junction and the measuring junction. The measurement may e made with an indicating or a recording potentiometer. Multiple- oint recording potentiometers are most useful when many oints must be read with ease and rapidity. The accuracy of such struments is sufficiently high for measuring air temperature. y regulating the size of the junctions or thermocouples, any peed of response desired may be obtained. When very fine junc- ons are used, little or no protection from radiation is needed if atural ventilation is not restricted. Very small thermocouples are, owever, very sensitive to fluctuations in temperature which may e a minor disadvantage. Another advantage in using thermo- ouples is that they may be adapted for the measurement of humid- y or wind speed which can also be read on the same indicating r recording potentiometer.

RESISTANCE THERMOMETERS

By passing a small current through a resistance which changes ith temperature, such as a thermistor, it is possible to determine ne temperature of the thermistor, using a resistance measuring evice. Generally, more accuracy is attainable with this method nan with thermocouples, but greater difficulties are usually en- ountered in calibration.

II. Humidity

The effect of the humidity of the air on man is often of consider- ble importance. To measure humidity, as well as to relate its ffect on man, the rather vague idea of "humidity of the air" nust be expressed in more specific terms. All expressions of humid- y of the air are direct or indirect expressions of the pressure of aseous water molecules known as water vapor in the air. The

movement of gaseous and liquid water molecules depends on temperature. At the boundary of a water and air interface, there is always an exchange of water molecules—some leaving the water surface and some entering the water. At higher temperatures more water molecules escape from the water surface than at lower temperatures, due to the increased velocity of the molecules, thus increasing the number of molecules of water vapor. At a given water surface, when more molecules are leaving than returning, evaporation is taking place. When a molecule escapes from the water surface, it does work and heat is lost, cooling the remaining water. When there is equilibrium, i.e., when the same number of molecules are entering and leaving the water surface, the pressure of the water vapor is at its maximum value, and at any particular temperature the equilibrium state is known as the *saturated vapor pressure*. If more molecules of water vapor are entering the water than leaving it, the saturated vapor pressure has been exceeded and condensation takes place. The saturated vapor pressure, for all temperatures above freezing, is almost solely dependent on the temperature, being higher at higher temperatures.

This same type of exchange takes place at the surface of ice, but the saturated vapor pressure over ice is slightly lower than that over supercooled water at the same temperature. At temperatures below freezing there are, therefore, 2 "saturated vapor pressures" for each temperature—one in relation to ice and the other in relation to supercooled water. When ice and supercooled water are in close proximity in an environment where there is a saturated vapor pressure with respect to water, it is supersaturated with respect to ice. Condensation (or sublimation) occurs on the ice and removes some of the molecules from the vapor stage. This lowers the vapor pressure and permits evaporation from the water. In practice, tables of saturated vapor pressure often refer to water above freezing and to ice below freezing. They also refer to a flat surface of water or ice. The vapor pressure is slightly higher over convex or irregular surfaces.

Vapor pressure is expressed as a partial pressure of the atmosphere and is expressed in the same units. These units may be inches of mercury, millimeters of mercury, or millibars.

Relative humidity is the actual water vapor pressure divided by the saturated vapor pressure for the temperature and is expressed as a percentage. Below freezing, 2 relative humidities can be found for each condition, depending on which saturated vapor pressure is chosen.

The *dew point* or dew point temperature is the temperature at which condensation occurs on a flat surface when the air is cooled.

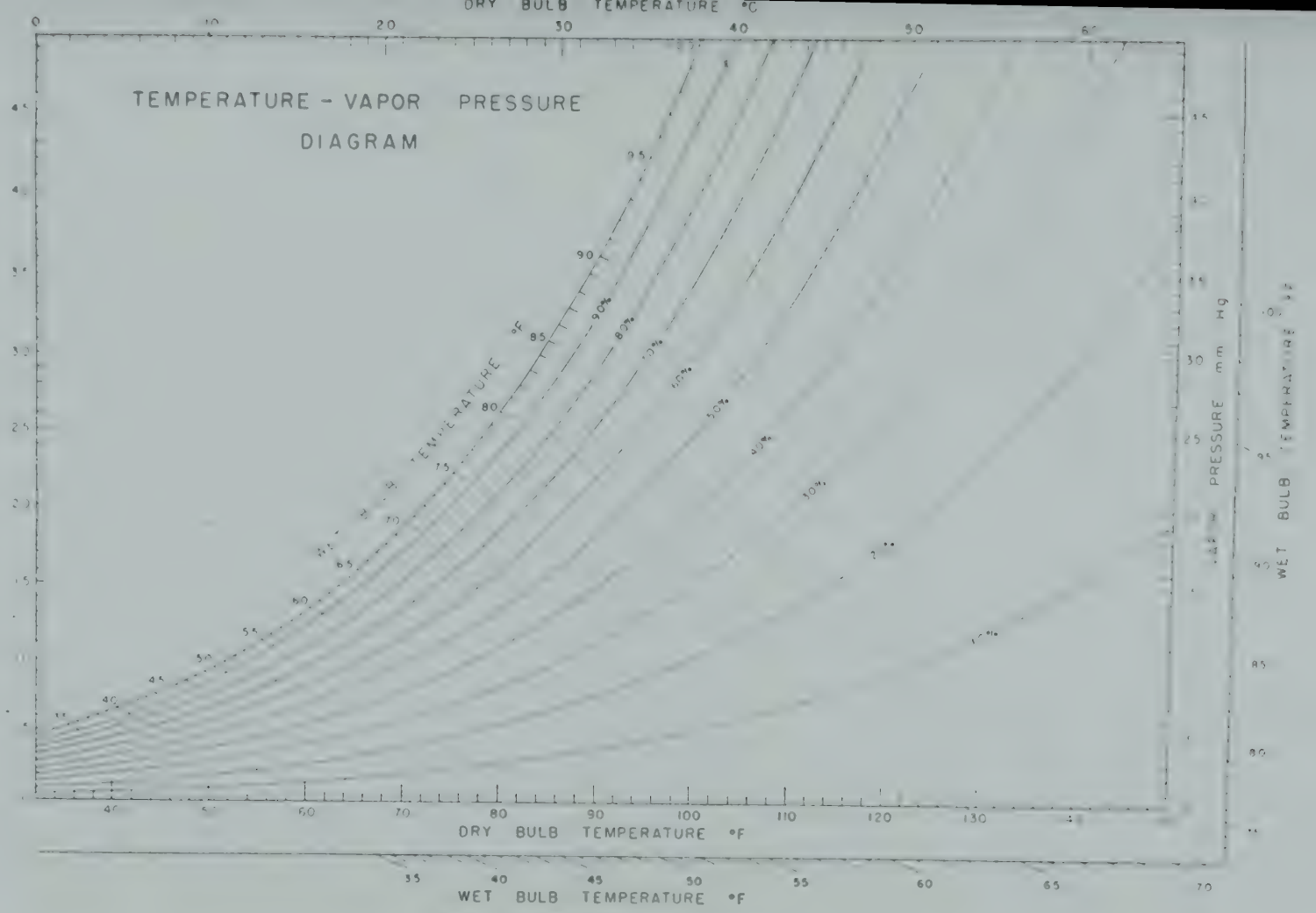


FIG. 1.—Psychrometric chart.

When the dew point temperature is the same as the air temperature, the saturated vapor pressure has been reached and the relative humidity is 100%. If the actual vapor pressure is known, the dew point may be calculated from the temperature at which the saturated vapor pressure is the same as the actual vapor pressure of the sample of air.

There are several other expressions of humidity that are less frequently used. These include *specific humidity*, which is expressed as g of water vapor/kg of air, including the water vapor in it, or in gr of water/lb of air. *Mixing ratio* is the same, except that dry air is used as the reference. *Absolute humidity* is expressed as weight of water vapor/unit volume of air.

Regardless of the method of measuring humidity, all of the above expressions may be derived from tables, formulas or from psychrometric charts. The psychrometric chart is usually the most convenient method and, if sufficiently large for readability, quite accurate. A small-scale reproduction of such a chart is presented as Figure 1.

PSYCHROMETERS

1. *Sling psychrometer.*—This is one of the most accurate, cheapest and most used instruments for measuring humidity. Two similar glass tube mercury-filled thermometers are attached to a metallic backing. A wick which covers 1 of the bulbs is moistened with clean or distilled water immediately before a reading is made. The top of the backing is attached to a swivel which in turn is attached to a handle. The whole assembly is then rapidly whirled to ventilate thoroughly both the wet and the dry thermometer bulbs. The temperature of each thermometer is read. The psychrometer is again whirled and another reading made. This is continued until the wet thermometer or "wet bulb" temperature has reached its lowest point. Both temperatures are then recorded. Vapor pressure, dew point, relative humidity, etc., may be determined from the temperatures through the use of psychrometric tables or a psychrometric chart.

2. *Ventilated psychrometers.*—There are a number of ventilated psychrometers which produce the same or better results and are usually easier to read. In these the thermometer remains still and air is drawn over the bulbs. This may be done with electric suction fans, hand-cranked fans, spring-driven fans or with suction produced by pumping a rubber bulb or bellows.

3. *Other psychrometers.*—It is not necessary to use mercury in glass thermometers; any pair of temperature-sensitive elements can be used if not damaged by contact with water. For example, ther-

thermocouples may be used with 1 covered by a wick to keep it wet. Such psychrometers, although they may be small and delicate, require little ventilation. Recording psychrometers can be made, using 2 temperature sensitive elements, 1 of which is kept wet by a wick and sufficiently ventilated.

Errors in psychrometers.—All the errors in psychrometers are likely to give too high an observed humidity, assuming the temperature-sensitive elements are accurate. Possible errors include stem conduction, dirty water or dirty wicks or too thick a wick, or too little ventilation. Sufficient ventilation depends on the size of the temperature-sensitive element and the size of the wick. For an average-sized mercury bulb, a ventilation of about 10 mph is ample and is easily accomplished. Very fine thermocouples may be sufficiently ventilated with less than 1 mph wind speed.

HYGROMETERS AND DEW POINT MEASURING INSTRUMENTS

Hygrometers measure relative humidity by the contraction and expansion of hygroscopic materials. The *hair hygrometer* is an example of this type of instrument. Suitably treated human hair is often used to actuate a pointer or a pen to indicate or record the relative humidity. These instruments are easy to use and may be employed if great accuracy is not necessary and if they are not used under certain conditions. They are subject to considerable error at low humidities and are not recommended for use in very low temperatures, where the time lag becomes very great. At −40 F, the lag is almost infinite. Other materials such as goldbeaters' skin are also used with considerable success as the hygroscopic sensitive element.

1. *Dew point hygrometers.*—These instruments may be extremely accurate but at the expense of considerable effort and cost. In essence, a polished surface is cooled until dew begins to form. At this point the temperature of the surface is measured and the dew point is determined. Actually, this method is difficult and is not particularly recommended for field use. Many much more simple instruments with sufficient accuracy are available.

*Dewcel.**—This is an example of another type of dew point measuring instrument. The operation of the sensitive element in this type of instrument has been described by the manufacturer:

"The basis of operation of the Dewcel is the behavior of a hygroscopic salt in the presence of water vapor. If dry lithium chloride is exposed to the atmosphere under average room conditions, it will

* Foxboro Company, Foxboro, Mass.

absorb moisture and dissolve, forming a saturated solution. If the salt and saturated solution is heated to higher temperatures, the water in the solution tends to escape to the atmosphere until a temperature is reached where the escaping tendency of the water is just equal to the tendency of the salt to take on water from the atmosphere. In other words, the vapor pressure exerted by the water in the solution is equal to the partial pressure of the water vapor in the atmosphere and an equilibrium condition exists. At this equilibrium point, the temperature of the salt and the saturated solution (temperature of the Dewcel) is a measure of the partial pressure of water vapor in the surrounding atmosphere, i.e., dew point temperature. The humidity sensitive Dewcel consists essentially of a thermometer bulb inside a thin-walled metal tube covered with a woven glass tape impregnated with lithium chloride. The tube is wound with a pair of silver wires over the tape and covered with a perforated metal guard. When the lithium chloride salt absorbs moisture from the surrounding atmosphere, it becomes an electrical conductor so that current passes between the two wires, thereby raising the temperature of the Dewcel until equilibrium is attained. This equilibrium temperature is measured with a conventional temperature bulb, either liquid expansion or electrical resistance type. The entire operation is simple, positive and completely automatic."

The temperature thus obtained may be indicated or recorded on a suitable recorder.

Discussion.—Humidity does not vary much between the level of the ground and 6 or 8 ft above the ground; consequently, the height of the measuring device is not as important as the height of the temperature-measuring device. Of course, the relative humidity, being dependent on temperature, will vary, but with 1 humidity observation the relative humidity may be determined for other levels at which the air temperature is known. For example, if at 6 ft the dry bulb is 90 F and the wet bulb 80 F, the vapor pressure is about 23.8 mm Hg as determined from a psychrometric chart. At 90 F the saturated vapor pressure is about 36.1 mm Hg, and at this level the relative humidity would be about 66%. If the temperature at 1 ft above the ground were 100 F and it were assumed that the vapor pressure was about the same (a reasonably safe assumption except when the surface is wet and rapid evaporation takes place), the saturated vapor pressure at 100 F is about 49.1 mm Hg and the relative humidity at this level would be about 48%. If it is assumed that the vapor pressure is the same at both levels, the dew point temperature would also be the same.

Barometric pressure should be taken into consideration, for there

a difference in psychrometric tables and charts due to pressure. Daily fluctuations in pressure are usually too small to bother about, but at different elevations tables and charts for the appropriate average pressure should be used.

III. Wind Speed

The standard practice in meteorologic observations of wind speed is to raise the measuring device high enough above the ground to obtain a representative wind for a large area and also to have the device far enough away from buildings, trees, etc., to minimize their influence in reducing wind speed. This procedure is, of course, unsatisfactory for the purpose of measuring the environment of man, who is usually on the ground and not on top of a 15 ft pole in a large field. Between a 15 ft level and the ground, the wind speed drops until at the lowest layer next to the ground the air film is almost calm at all times. Computations have been made to determine the decrease in wind speed as the ground is approached and are, no doubt, reasonably accurate over flat surfaces such as water, short grass, etc. The complexities are so great, however, in attempting to calculate all of the variables due to uneven ground, different lapse rates (rate of temperature change with altitude), obstructions such as bushes, trees, etc., that by far the easiest method to determine the wind speed of the environment of man is to forget about the standard wind speed measurements and measure it at the level of, and in the vicinity of, the subject himself. The speed of the wind or the ventilation is often an important part of the environment of a man. Usually, it is the cooling or heating power of the wind, or the speed of evaporation of sweat, which is ultimately of greater interest than how many miles per hour the wind is blowing. If this is the case, an anemometer which measures the cooling power of the wind is often better suited than 1 of more conventional design which records air movement.

WINDMILL ANEMOMETERS OR AIR METERS

Windmill anemometers which are held by hand are quite accurate even at relatively low wind speeds if they are constantly held facing the wind. In a wind tunnel this may be possible, but in the field it is very difficult because near the ground the wind is apt to be gusty and change direction rapidly. They are, therefore, not suitable for outdoor measurements unless attached to a wind vane which will faithfully keep them pointing into the wind. Even when this is done, difficulties are encountered in reading the instrument because the dials indicating the amount of wind which

has passed are on the front of the instrument itself. A non-directional anemometer will be found more desirable in most cases. If no other instrument is available, a lightweight thread about 1 ft long may be tied to the instrument. This will blow away from the wind and will act as a guide for aiming the windmill correctly by hand.

2. CUP ANEMOMETERS

Cup type anemometers have a great advantage over the hand-held windmill type since they are non-directional. This is the type most frequently used for meteorologic observations. It is less accurate than the windmill type at low wind speeds (some may not even start until the wind is 2 or 3 mph). The older types generally have 4 cups, but the newer have 3 cups and are more accurate. Some measure "miles of wind passed" on a dial at the base of the anemometer; others, which turn an electric generator, may be read remotely by means of a voltmeter calibrated in miles per hour, knots, kilometers, etc. Many of the first type have an electrical contact which closes when each $^1/_{60}$ of a mile of "wind" has passed and another contact for each mile which passes. If a buzzer or light is included in an electrical circuit with the $^1/_{60}$ mile contact, it is necessary only to count the number of contacts in 1 min to determine the wind speed of the last minute. A suitable recorder may be used to count the contacts.

3. SPECIAL CUP TYPE ANEMOMETER

A special cup type anemometer is being manufactured which is extremely sensitive to very low wind speeds and is capable of recording wind gustiness. Very light-weight cups are mounted on almost frictionless bearings. No gears or generator are used which require the spindle to do work; instead there is a hole or holes in the spindle through which a light shines on a photoelectric cell. As the spindle turns, the photoelectric cell receives pulses of light as the beam passes through the spindle holes. The impulses are relayed to an electronic device which actuates a pen on a moving chart calibrated in wind speed. This is an excellent instrument, especially for those who are interested in recording gustiness. For those desiring average wind speeds the instrument is a bit too sensitive. The cost is even higher than for the propeller type anemometer.

4. PROPELLER-TYPE ANEMOMETERS

Light-weight propeller type anemometers have recently been made which are mounted on the front of what resembles a small airplane without wings. The tail of the "airplane" keeps the propeller facing the wind. The propeller operates a small generator

and the power produced operates a voltmeter calibrated in wind speed or a recording voltmeter similarly calibrated. Greater accuracy is reported for this type of anemometer than with the cup type. Since they are easily mounted, fairly sensitive to low wind speeds and read remotely, they are supplanting the cup type for many purposes. This type of anemometer is, however, relatively expensive.

5. *PRESSURE PLATE ANEMOMETERS*

The pressure exerted on a flat plate held normal to the wind is roughly proportional to the square of the wind speed. Corrections for size of plate, density of the air, etc., must also be made. The plate may be held normal to the wind by means of a vane and the pressure applied by the wind may be measured in a number of ways. One which has been used is a plate hinged and attached at the bottom and held almost rigid with a heavy spring. As the pressure increases, the plate is pushed backward slightly and moves the core of a coil where impedance is changed. The change of impedance is measured and calibrated as wind speed. Another method could be one in which a strain gauge is used. In this case a slight deformation of the gauge can be measured because the strain gauge changes resistance with very slight deformation. Pressure plate anemometers are very sensitive and their response very rapid, thus recording gustiness rather than average wind speed over a period of time.

6. *HEATED THERMOMETER ANEMOMETERS*

Heated thermometer anemometers measure the cooling power of the wind which is related to the wind speed. One of the earlier types is the Katathermometer, developed by the physiologist Leonard Hill. In this type a liquid in a glass thermometer with a large bulb is heated to over 40 C, and the time taken for the temperature to drop from 38 C to 35 C is measured with a stop watch. By using a formula which includes the temperature of the air, instrumental constants and time of the 3° C drop, the average wind speed passing over the bulb of the thermometer for the period of measurement can be calculated. A more recent type utilizing the cooling power of the wind is made by electrically heating the bulb of a liquid in glass thermometer with a fixed electrical heat input. The temperature of the heated thermometer is read; the difference in temperature between the heated and an unheated thermometer is determined and the wind speed calculated from a calibrated curve which relates temperature difference to wind speed.

Work has been initiated at the Quartermaster Climatic Research

Laboratory to construct a heated thermometer anemometer utilizing thermocouples rather than a thermometer, and initial experiments show promising results. When completed, remote recording of the wind speed may be obtained, using a recording potentiometer.

All anemometers utilizing the cooling power of wind have the common disadvantage that they may not be used during any type of precipitation.

IV. Radiation

The radiant energy exchange between man and his environment can become quite complicated when one considers that a man may be losing heat by radiation to some objects and at the same time be gaining heat from other objects. Some knowledge of the different types of radiant energy exchange is necessary in order to understand the net radiation exchange between man and his environment.

Every surface emits radiant energy which is actually electromagnetic waves like radio waves (very long waves), light waves (short waves) or x-rays (very short waves). Radiant energy which most affects man under natural environmental conditions is made up of electromagnetic waves with wavelengths near those of light (short waves) and of infra-red waves (long waves). While every surface emits radiant energy, every surface will also receive or absorb radiant energy which on absorption is transformed to heat (radiant energy which is emitted removes heat from the surface). If one considers radiant energy in the region of visible light, then white surfaces are those which reflect most of the energy, while those which are black absorb most of the energy. In general, there is a law which says that if a surface is a good absorber in a certain wavelength, it is also a good emitter of radiation of the same wavelength. "Black body" is a term used for a 100% absorber and 100% emitter. The total amount of radiation emitted by a black body is proportional to the fourth power of its absolute temperature. The net exchange between 2 parallel surfaces is the difference between that emitted by the first and absorbed by the second and that emitted by the second and absorbed by the first. If both surfaces are the same temperature and both are "black bodies," the net transfer of radiant energy is zero because both are emitting and absorbing the same amount.

1. RADIATION THERMOPILE

Radiation, other than direct solar radiation, is generally meas-

red by comparing the relative temperature of 2 surfaces, 1 of hich is exposed to radiant energy exchange and the other shielded rom it by a cover and maintained at a known temperature. The hielded surface is often in contact with a relatively heavy block f metal with high specific heat so that the surface changes tem- erature slowly. The surface exposed to radiation is light in weight nd protected from ventilation; consequently, the exchange in adiant energy is more effective than the surrounding air in letermining the temperature of the surface (the opposite of the roper method for exposure of a thermometer when measuring air emperature). In order to measure the difference in temperature, hich may be very slight, a series of thermocouples is used which nultiply the electromotive force produced by a single thermo- ouple. Such a series of thermocouples is known as a thermopile. The surface exposed to radiation does not necessarily take on the emperature of the object toward which it is pointed due to the nodifying influence of the air around it and conduction within the nstrument itself. However, after the instrument has been cali- rated, a given electromotive force as measured on a suitable otentiometer may be converted to units of heat exchange even hough the relationship is not linear. Usually, the heat exchange is xpressed in langleys (g-cal/min/cm^2) but may be converted to g-cal/m^2/hr by multiplying by 600. If the temperature of the urface toward which the instrument is pointed is desired, rather han the energy exchange between the surface and the instrument, t may be determined from the formula

$$R = C(T_i^4 - T_s^4)$$

where R is in units of heat exchange, C a constant, T_i the absolute emperature of the instrument (the temperature of the block is sually used) and T_s the absolute temperature of the surface meas- red. The constant used depends on the units of heat exchange sed. If the units are kg-cal/m^2/hr, the constant to be used is 4.92×10^{-8}.

Some radiation thermopile instruments are supplied with poten- iometers calibrated in temperature difference rather than in millivolts. The above formula may then be used to determine the heat exchange units.

The determination of the long wave radiation exchange between man and his environment is complicated by the fact that usually the effective black body temperature is different in different directions. For example, the night sky is "cooler" overhead than near the horizon because there is less atmosphere (including dust and water vapor which radiate more or less like black bodies)

straight through it than on the diagonal. The ground may be still another temperature.

One method of determining the heating or cooling effect of long wave radiation on man is to measure radiation exchange in many representative directions, take the average and assume that the net radiant energy exchange is between man and this temperature. This method is not absolutely correct even if a perfect average were known since no consideration is given to the fact that radiation is proportional to the fourth power of the absolute temperature. In making the actual measurements, it is the difference in black body temperature of the environment (as averaged) and the temperature of the man (as measured by the radiation thermopile) which is desired rather than the temperature of the environment and the instrument.

During the daytime another complication arises because the instrument will not differentiate between long and short wave radiation unless a filter is used. Therefore, during daylight 2 sets of measurements must be taken, 1 without a filter and 1 with an infra-red filter, to make the instrument insensible to long wave radiation. To determine the long wave radiation, the difference between the measurements indicates the long wave radiation. In both sets of measurements, the instruments should not be pointed to the direct rays of the sun.

2. PYRHELIOMETERS

Most instruments used to measure solar radiation are known as pyrheliometers, though some are known by other names. A radiation thermopile may be used to measure solar radiation if heavily filtered, but most solar radiation measuring instruments are constructed slightly differently. Since most of the energy of solar radiation lies in the short wave and since most objects are not "black bodies" to short wave radiation, it is possible to construct an instrument which is almost independent of the temperature of the ambient air.

A black surface and a white surface, usually covered by a glass bulb to protect them from the wind, are exposed to the incident solar radiation. The black surface, absorbing more radiant energy, becomes warmer than the white surface. The difference in temperature is measured, usually with a thermopile, and the resulting electromotive force measured on a suitable instrument. The instrument, whether it be a potentiometer or a microvoltmeter, may be calibrated in units of heat exchange. Usually, the units used are langleys.

At meteorologic observatories the pyrheliometer is usually

mounted on the horizontal, though measurements may be made normal to the sun or in any other way desired. For measurement of the effect of short wave radiant energy exchange, special consideration must be given to the area of the man receiving radiation, the reflectivity of his skin and clothing, the effect of penetration of short wave radiation into his clothing, etc. It might be noted that the measurement of incident solar radiant energy is very simple, but its heating effect on man is a very complicated problem.

Comment by Ralph J. Wedgwood

This description of the principles and methods of measurement of the climate surrounding a man should be of prime interest to research workers and clinicians interested in accurate observations. It presents in simple and explicit language the materials and methods required for the accurate measurement of the weather conditions impinging on man. It should not have to be emphasized that these conditions in which man exists are becoming of increasing interest and importance to all forms of medical research.

REFERENCES

1. Blair, T. A. (ed.): *Weather Elements* (New York: Prentice-Hall, Inc., 1942).
2. Catalogs of the American Instrument Company, Silver Spring, Md.
3. Catalogs of the Eppley Laboratory, Newport, R. I.
4. Catalogs of Friez Instrument Division of the Bendix Aviation Corporation, Baltimore.
5. Foxboro Company Bulletin 407-K (Foxboro, Mass.).
6. Geiger, R. (ed.): *The Climate near the Ground*, tr. by Melroy M. Stewart (Cambridge, Mass.: Harvard University Press, 1950).
7. Hansmann, E., and E. P. Slack (ed.): *Physics* (New York: D. Van Nostrand Company, 1943).
8. Knowles, W. E. (ed.): *Meteorological Instruments* (Toronto: University of Toronto Press, 1947).

RESPIRATORY EXCHANGE*

LOREN D. CARLSON, *University of Washington*

MEASUREMENTS of respiratory exchange are usually made to determine metabolic activity by the indirect method. In addition, respiratory heat exchange enters into calculations of heat loss. It is the purpose here to discuss the methods for making measurements that may be feasible under field conditions and to discuss the various techniques involved.

Before any measuring technique is used, the limits of its accuracy must be assessed in terms of the desired tolerance. If the respiratory exchange is to be used in calculating energy output or metabolic rate, inspired or expired volume and oxygen content of expired air are determined. With these measurements and an assumed caloric value for a liter of oxygen, the calculation can be made. On some occasions, a total respiratory quotient is used to verify the caloric value assigned to a liter of oxygen. Thus, determination of respiratory exchange requires measurement of inspired or expired volume and of the concentration of carbon dioxide and oxygen in both the inspired and expired gas. Additional factors such as pressure, temperature and water vapor in the gas may also influence these calculations.

I. Calculation of Respiratory Exchange

The parameters involved are indicated in the formulation adapted from Weir (19). The formulation applies when the inspired oxygen fraction is 0.2093 and inspired carbon dioxide is negligible.

$$V_{O_2} = x + y + z = 1 \qquad \dagger(1)$$

* The work reported here was supported in part by contract between the University of Washington and the Alaskan Air Command, Arctic Aeromedical Laboratory (Contract AF 33 (038)-422), Ladd Air Force Base, Fairbanks, Alaska.

† Symbols for equations (1) through (9):

V_{O_2} = volume oxygen consumed/unit time
V_{CO_2} = volume carbon dioxide evolved/unit time
x = oxygen used to oxidize carbohydrate
y = oxygen used to oxidize protein
z = oxygen used to oxidize fat
K = kilocalories/unit time/liter of oxygen
R = respiratory quotient
P = fraction of protein contributing to caloric production

All volumes expressed dry at 760 mm Hg, 0°C.

$$V_{CO_2} = R = x + 0.802y + 0.718z \quad (2)$$

$$K = 5.047x + 4.463y + 4.735z \quad (3)$$

Substituting for x in equations (2) and (3)

$$R = 1 - 0.198y - 0.282z \quad (4)$$

$$K = 5.047 - 0.584y - 0.312z \quad (5)$$

then

$$K = 3.941 + 1.106R - 0.365y \quad (6)$$

Here, y represents the error when protein metabolism is not considered. If y equals zero, R becomes the nonprotein RQ. Since 1 g of urinary nitrogen equals 5.941 liters of oxygen (13), the correction is $0.365 \times 5.941 = 2.17$ kg-cal g urinary nitrogen. Alternatively, if the fraction of total caloric production due to protein is P, the number of kilogram-calories produced by protein metabolism per liter of total oxygen consumed is PK. The volume of oxygen then consumed becomes $\frac{PK}{4.463}$ liters, and equation (6) can be written

$$K = 3.941 + 1.106R - 0.365 \frac{PK}{4.463} \quad (7)$$

and simplified

$$K = \frac{3.941 + 1.106\,R}{1 + 0.082\,P} \quad (8)$$

Thus, as Weir (19) pointed out, the protein correction is equal to a deduction of 1% when $P = 0.123$ (12%). In man, P is usually 1/8, and equation (8) reduces to approximately

$$\text{Kilocalories} = 3.9\,V_{O_2} + 1.1\,V_{CO_2} \quad (9)$$

Weir carried his analysis further to determine the caloric value of a liter of expired air.

$$F_{O_2} = \frac{F_{EN_2}F_{IO_2} - F_{EO_2}F_{IN_2}}{F_{IN_2}} \quad \ddagger(10)$$

$$F_{CO_2} = \frac{F_{ECO_2}F_{IN_2} - F_{ICO_2}F_{EN_2}}{F_{IN_2}} \quad (11)$$

‡ Symbols for equations (10) through (19):

F_{O_2} = fraction of oxygen taken up/liter of expired gas
F_{CO_2} = fraction of carbon dioxide evolved/liter of expired gas
F_{IO_2} = fraction of inspired oxygen
F_{ICO_2} = fraction of inspired carbon dioxide
F_{IN_2} = fraction of inspired nitrogen
F_{EO_2} = fraction of expired oxygen
F_{ECO_2} = fraction of expired carbon dioxide
F_{EN_2} = fraction of expired nitrogen
K' = kilocalories/liter of expired air

All volumes expressed dry at 760 mm Hg, 0°C.

Substituting for F_{O_2} and F_{CO_2} and eliminating F_{EN_2} by putting $F_{EN_2} = 1.00 - F_{ECO_2} - F_{EO_2}$ makes equation (8) become

$$K' = \frac{1.0432 - 4.984\,F_{EO_2} + 0.063\,F_{ECO_2}}{1 + 0.082\,P} \tag{12}$$

The small coefficient of F_{ECO_2} indicates that K' is practically independent of the fraction of carbon dioxide.

One liter of expired air contains F_{EO_2} liters of oxygen. The corresponding volume of inspired air is $1 + F_{O_2} - F_{CO_2} = 1 + (1 - R)\,F_{O_2}$ liters. This contains $[1 + (1 - R)F_{O_2}]F_{IO_2}$ liters of oxygen. Then

$$F_{O_2} = \frac{F_{IO_2} - F_{EO_2}}{1 - (1 - R)\,F_{IO_2}} \tag{13}$$

or

$$F_{O_2} = \frac{0.2093 - F_{EO_2}}{0.7907 - 0.2093\,R} \tag{14}$$

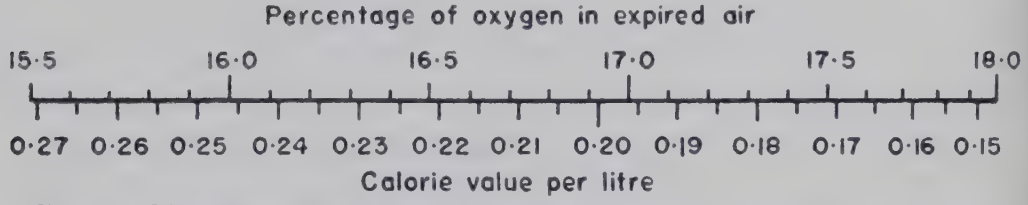


FIG. 1.—Nomogram giving caloric value/liter of expired air from percentage of O_2 in expired air. Heat output is obtained by multiplying number of liters of expired air by caloric value/liter. A protein correction is included on the assumption that 10–15% of the total calories arises from protein metabolism. (From Weir (19).)

Multiplying by the caloric value per liter

$$K' = \frac{(0.2093 - F_{EO_2})(3.941 + 1.106\,R)}{0.7907 + 0.2093\,R} \tag{15}$$

For range R, 0.718 to 1.0, K' lies between

$$5.032(0.2093 - F_{EO_2}) \text{ and } 5.047(0.2093 - F_{EO_2}) \tag{16}$$

so with an error of less than 1 in 600

$$K' = 5.04(0.2093 - F_{EO_2}) \tag{17}$$

and including a protein correction ($12^1/_2\%$ calories from protein)

$$K' = \frac{0.0504(0.2093 - F_{EO_2})}{1 + 0.082\,P} \tag{18}$$

$$K' = 1.046 - 5F_{EO_2} \tag{19}$$

The value of K' can be taken from Weir's nomogram (Fig. 1).

Note that this applies only when F_{IO_2} is 0.2093. If F_{IO_2} approaches 1, the error introduced by variations in R becomes significantly large.

With the greatest protein variation, this approximation of a caloric value for a liter of expired air determined by the expired F_{O_2} will be in error by $\pm 1\%$. Therefore, all measurements of volume and gas concentrations should be this accurate. Thus, an expired minute volume of 10 liters/min should be measured to the nearest 0.1 liter and an expired oxygen fraction in the neighborhood of 0.150, to the nearest 0.002. Since respiration is periodic, minute volume should be measured over a sufficient number of breaths to minimize the error.

The calculations are made with all volumes and fractions expressed STPD, so the temperature of the expired air volume must be determined along with the saturation of the gas. If the temperature of gas collection is below body temperature, the gas is usually assumed to be saturated and the temperature should be recorded to the nearest degree Centigrade. Nomograms which give the factor for finding the dry volume of a gas saturated with water vapor are available (7, 19).

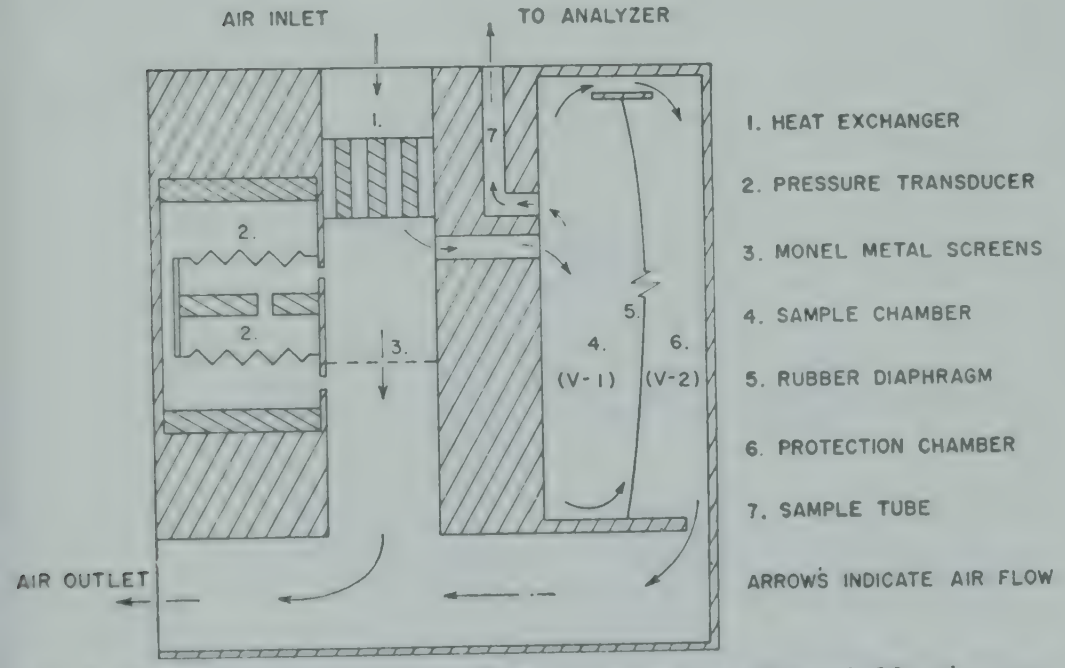


FIG. 2.—Schematic representation of flowmeter, using 400-mesh Monel screen and sampling chamber. (From Young *et al.* (22).)

II. Measurement of Respiratory Volume

Consolazio and his co-workers (7) have described in great detail the use of spirometers and Douglas bags to collect expired gas.

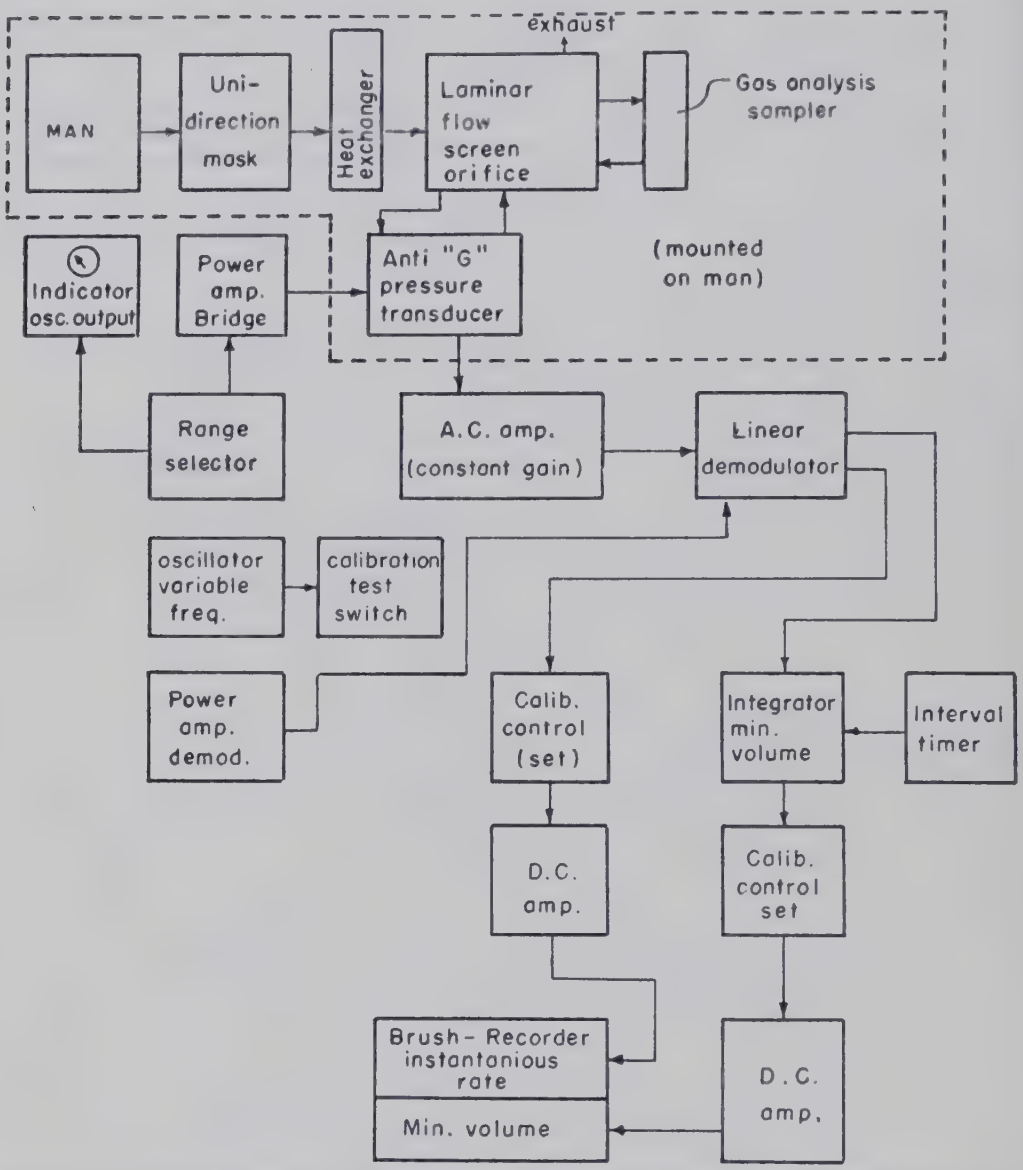


FIG. 3.—Block diagram of flowmeter and recording assembly. (From Young *et al.* (22).)

The usefulness of the latter is restricted by their bulkiness and relatively limited capacity (maximum, 200 liters).

The development of flowmeters which measure instantaneous rates of flow has introduced another means of measuring volumes. If a flowmeter is so constructed that the transducer output is

inear with flow, the flow may be integrated to give volume. 'olume 2 of *Methods in Medical Research* contains two descrip- ions (11, 15) of this type of flowmeter.

Young (21, 23) has developed this principle in 2 flowmeters vhich, with their recording apparatus, are capable of recording nstantaneous flow, average flow or minute volume. In addition, gas sampling is simplified. (The addition of a conventional count- ng circuit to the instrument would give respiratory rate as well.)

PORTABLE MONEL MESH FLOWMETER

This flowmeter (Fig. 2), connected to the recording apparatus by a cable and gas sampling tube (23), is on the expiratory side of a unidirectional mask. The expired air flows through a heat exchanger and the laminar-flow, screen orifice. The temperature of the flowmeter, which is covered with suitable heated insulation, s maintained at 39° C to avoid condensation in the unit and to acilitate flow and volume determinations. The pressure drop across two 400-mesh Monel screens is transmitted to a pressure ransducer which consists of a balanced condenser bridge con- tructed with 2 movable diaphragms. The block diagram of the electronic components (Fig. 3) indicates the parts. The original description (23) should be consulted for details.

CONCENTRIC CYLINDER TYPE FLOWMETER

The portable flowmeter using Monel screens has the disadvan- age of limited flow within the linear range and is easily con- aminated. To extend the range of the flowmeter and avoid the contamination, Young (21) developed the device shown in Figure 4. n this flowmeter, the laminar flow occurs between 3 concentric cylinders with the gaps between the shells set at $^1/_{32}$ in. The diameter of the shells (and their length) is determined by the maximum rate of flow expected. A diameter of 6 in. was found necessary to insure a pressure drop proportional to flow at rates up to 300 liters min. The pressure drop, which is approximately in. of water at this flow, is measured by a capacity type trans- ducer which, with the associated bridge circuit, develops a voltage proportional to the pressure drop.

Comment.—As a general rule, the research worker does well to use a direct volume device when possible, resorting to transducer systems only when speed of response or convenience requires them.

Recent reports of Glasow and Müller (Arbeitsphysiol. 14: 319–

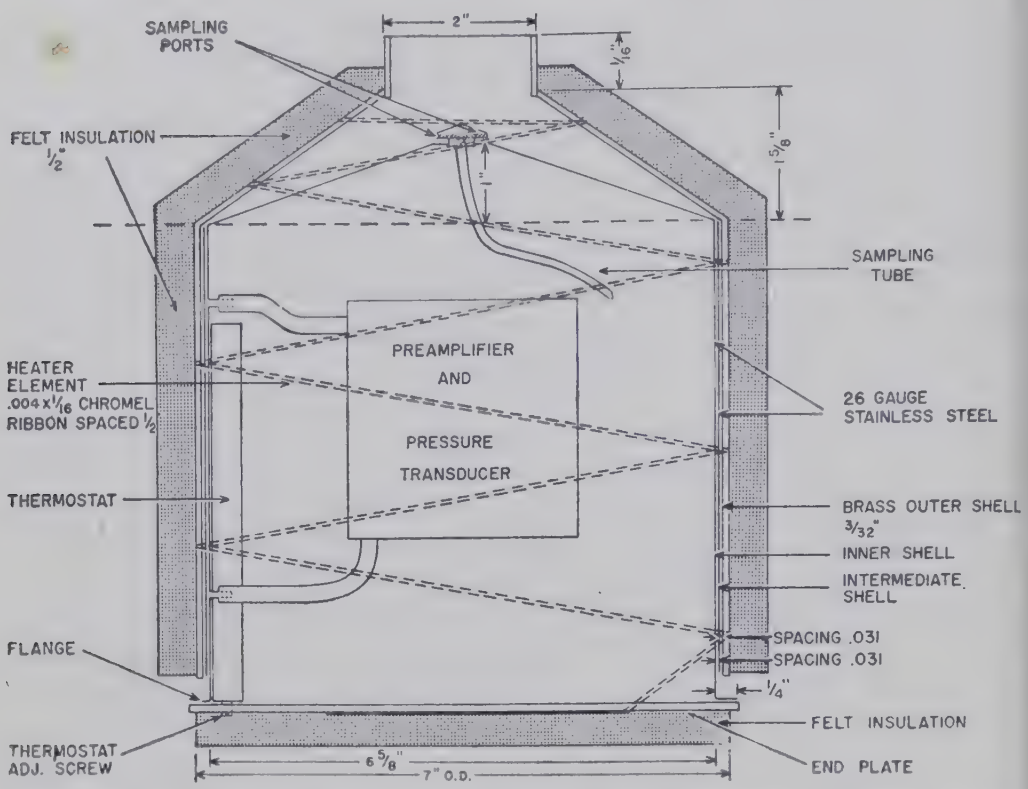


FIG. 4.—Basic unit of concentric cylinder type flowmeter. (From Young (21).)

321 and 322–327, 1951) call attention to the portable apparatus for determining gas exchange described by Kofranyi and Michaelis (see comment, this volume, p. 78).

III. Measurement of Gas Concentration

Once the expired volume is determined, the gas concentration must be measured. The expired gas consists of a dead air volume and an "alveolar" volume which must be mixed. In the Douglas bag or spirometer, adequate mixing can be obtained and measures of gas concentration in expired air easily made. However, in the gasmeter or flowmeter, the total sample is not collected, and the sampling procedure must accurately reflect the gas concentration in mixed expired air. The sample cannot be continuously drawn by pump since air will be drawn back during the inspiratory cycle. The physical principle of the flowmeter is again utilized to advantage by allowing the pressure drop in the flow-measuring device to force air through the sampling or analyzing chamber (e.g., Pauling meter).

An ingenious sampling chamber has been described by Young *et al.* (22). Expired air is forced into the chamber *V–1* (Fig. 2) by the pressure developed in the flowmeter. This distends the rubber

diaphragm sufficiently to maintain a positive pressure in *V–1* during inspiration. Since the pressure developed in the flowmeter is proportional to the rate of flow of expired air and the flow of gas into chamber *V–1* is proportional to the pressure at the inlet, the flow of gas into the sampling chamber is proportional to the flow of expired air. The function of chamber *V–2* is to prevent outside air entering *V–1* due to movement. The upper limit of operation is determined by the effective volume of chamber *V–1*. A continuous sample for analysis can be withdrawn at a slower rate of flow (1 cc/sec).

If determination of metabolic rate is the purpose of respiratory exchange measurements, oxygen concentration in expired air is another parameter. In certain cases, it is desirable to measure

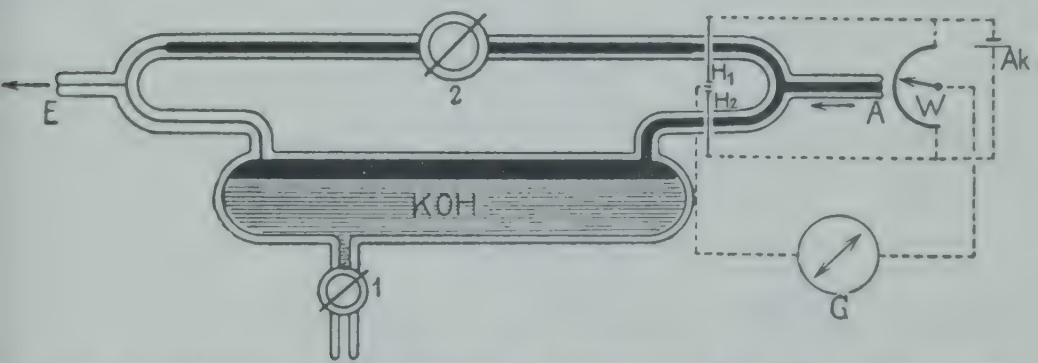


FIG. 5.—Schematic representation of apparatus for CO_2 determination. *A*, entrance for gas; *E*, exit of gas; H_1 and H_2, hot wire flowmeter; *W*, variable resistance; *G*, galvanometer; *AK*, battery. (From Rein (14).)

carbon dioxide concentrations as well. Chemical methods of analyzing for carbon dioxide and oxygen have been described in detail (7). Particular notice should be taken of the Scholander micrometer gas analyzer which permits determination of oxygen, carbon dioxide and nitrogen in 0.5 ml samples with an accuracy of ±0.015%. This analysis requires 6–8 min. Also deserving of consideration is the Fry analyzer (10), which is economical and simple to handle. It analyzes a 2–3 ml sample with an accuracy of ±0.5%.

Rein (14) combined the chemical method with a flow-sensitive device to obtain continuous recording of gas concentration (Fig. 5). If a component of a gas flowing through 2 chambers is absorbed in 1 chamber, the rate of flow becomes imbalanced. The Rein analyzer measures this imbalance.

There are several convenient and accurate physical methods for gas analysis. For oxygen measurement, the Beckman Model C oxygen analyzer has become standard in many laboratories. Its

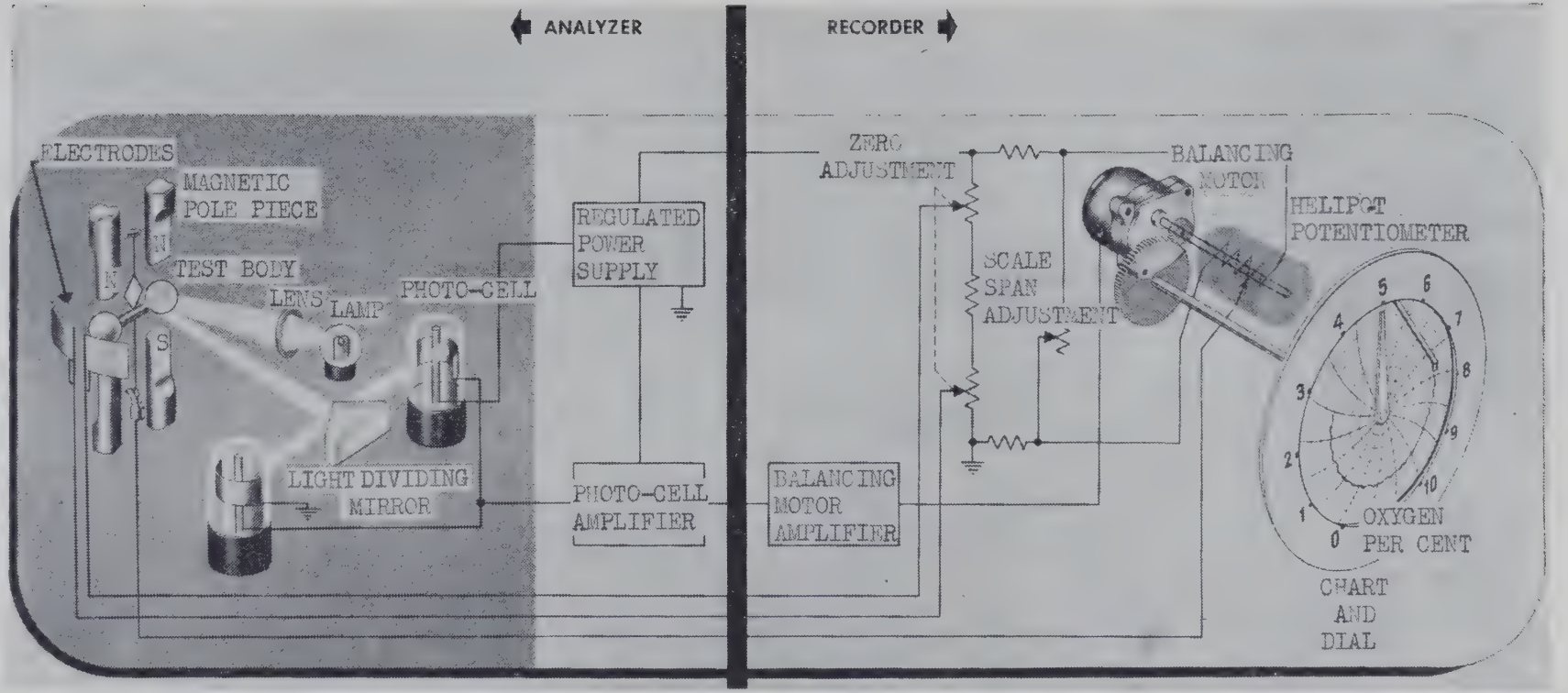


FIG. 6.—Operation of Model G-2 Beckman oxygen analyzer. A beam of light reflected from the mirror attached to test body is deflected in accordance with any rotation of test body due to presence of a magnetic sample gas. The light beam is divided between 2 photocells. Any motion of the test body due to changes in oxygen concentration results in an unbalance in the light on the 2 photocells. This unbalance produces an electrical current which is amplified by the photo-cell amplifier in the analyzer and balancing motor amplifier in the recorder. This current is used to drive a balancing motor forward or backward depending upon the direction of rotation of the test body. The motor is geared to the shaft of a Helipot potentiometer which supplies a variable electrostatic potential to the test body. By this mechanism, the potential required to hold the test body in null position is automatically and continuously supplied. Any change in magnetic force due to a change in O_2 partial pressure is always balanced by an electrostatic force which, in turn, is proportional to d-c voltage delivered by the potentiometer. Thus, changes in O_2 partial pressure are accurately translated into proportional voltage changes which can easily be recorded on commercially available recording potentiometers. (Courtesy Arnold O. Beckman, Inc.)

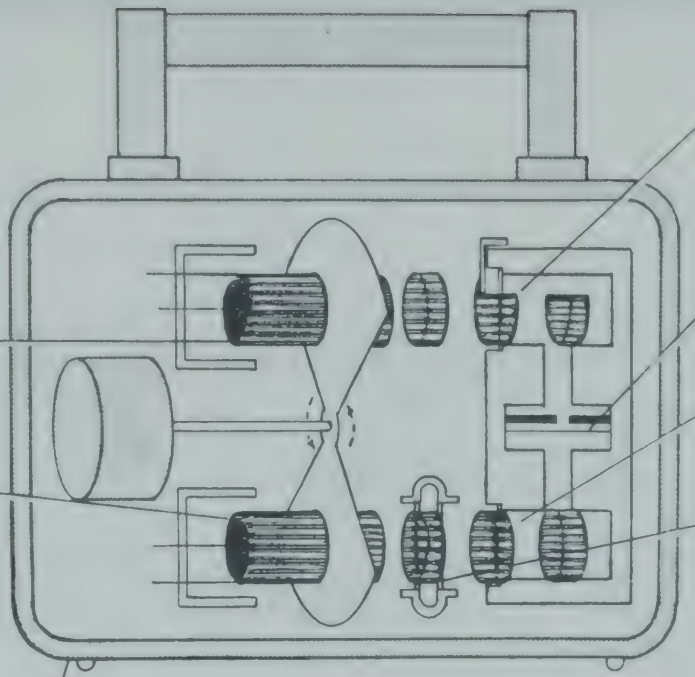


FIG. 7.—Principle of infra-red measurement in CO_2 analyzer. (Courtesy Liston Becker Instrument Company)

operation depends on the paramagnetic properties of oxygen. This instrument can be obtained with a full scale range of 60 mm or 760 mm with an accuracy of ±1%. The response time is of the order of 10 sec if the gas sample is passed through the pole pieces, requiring close flow rate control, and approximately 50 sec if the analyzer is protected by a fine porous diffusion plate. A recording analyzer, shown schematically in Figure 6, is obtainable. Young and Jones (23) described an oxygen analyzer, utilizing the paramagnetic effect, with a response (70%) time of 0.07 sec which is suitable for use in respiratory measurements.

For carbon dioxide measurement, conductivity, light refraction and infra-red absorption methods have been used. The Cambridge Instrument Company manufactures a thermal conductivity cell.

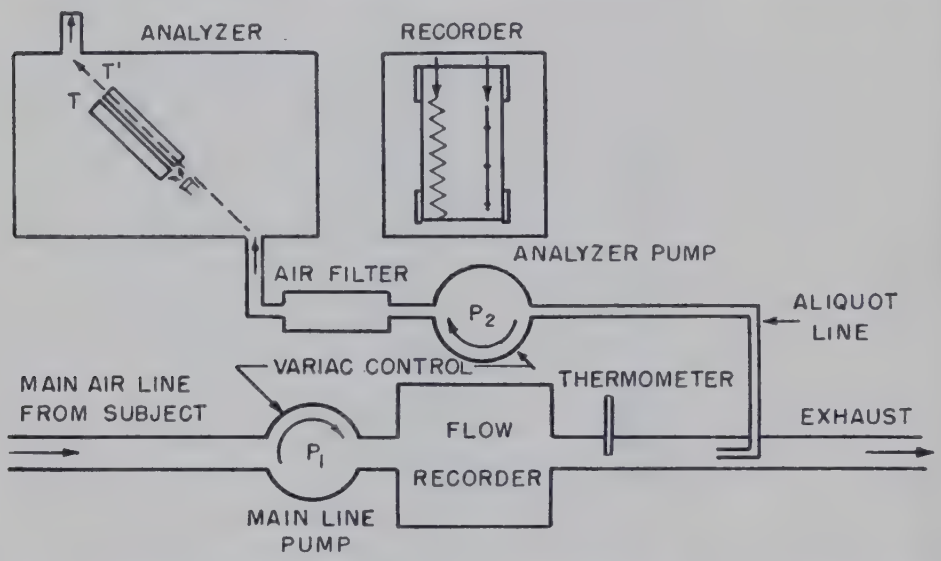


FIG. 8.—Schematic diagram of air pumping system and analyzing system used by Spoor (16).

The refractometer has been described by Clamans (6). A number of infra-red analyzers are described in the literature (4, 9, 12, 20), and the Liston Becker Company manufactures a commercial model. (Baird Instrument Co. and Leeds and Northrup manufacture instruments that are not as suitable.) Figure 7 shows a schematic diagram of the Liston Becker instrument and adequately describes its operation. Two models are made. The range may be as small as 0–3% or as large as 0–30% carbon dioxide. Accuracy is ±0.1% carbon dioxide, and speed of response (90%) is 0.1 or 0.3 sec, depending on the model.

Spoor (16), in addition to describing the use of a Leeds and Northrup infra-red analyzer, reported a method of determining

metabolism which differs from the methods described above (see Fig. 8). A constant volume of air exceeding in amount the maximal respiratory volume is drawn past the subject by a pumping system. An analyzer aliquot is drawn off by means of a small constant speed rotary pump. Other investigators using this principle (e.g., Benzinger) introduce a rubber bag of 2–3 liter capacity between the subject and the measuring devices to "buffer" the peaks in flow velocity.

IV. Temperature and Saturation of Expired Air

A common assumption made in respiratory gas exchange calculations is that the gas is saturated at 37 C. Although it has been known for some time that this is not a valid assumption even in temperate climates and particularly not in cold environments (17, 18), the value is used because this variable is difficult to measure. Webb (17, 18) measured temperature in the respiratory tract with a fine thermocouple. The flowmeter of Young (20, 22) keeps the gas above 37 C, making the water vapor correction the only assumed value. In his instrument, to introduce a 1% error, the expired air need contain 40 mm Hg water vapor instead of 47 mm Hg. This is the saturation pressure at 34 C. For accurate respiratory measurements it would be suitable to measure the temperature and the saturation of the expired air. Such saturation measurements can be made with American Instrument Company's Dunmore hygrometers.

There may be occasions when alveolar air measurements are desired. The techniques used in this analysis can be found in Volume 2 of *Methods in Medical Research* and in the reports of Du Bois *et al.* (8) and Young (20).

V. Estimation of Metabolic Effort

For an approximation of metabolic effort in the field, minute volume could suffice since it varies linearly with oxygen uptake. There is great need for a simple ventilation meter for field tests similar to one produced for clinical measurements by Bennett (Respiratory Ventilator Meter Model RVM, J. J. Monoghan Co., Denver). Berggren and Christensen's (3) demonstration of the relationship between the heart rate and oxygen consumption during certain types of work offers another possible method for establishing an approximation of work loads.

NOTE.—This section was reviewed by Allan Young.

Comment by Ulrich C. Luft

The mathematical analysis of the caloric equivalents for oxygen "in vivo" well demonstrates that it is permissible for most purposes to forego a quantitative correction for protein metabolism in indirect calorimetry. The more comprehensive procedure involving the determination of urinary nitrogen will be limited to instances where specific information regarding the participating energy substrates is desired.

Both flowmeters designed by Young, Carlson and their associates represent important improvements of the pneumotachograph principle. When only the minute volume of respiration is required, dry gas meters* have also proved satisfactory for recording inspired volume by electric transmission of dial movements. Mean flow velocities of 50–60 liters/min can be attained without excessive resistance.

The interferometer (2, 5) also deserves consideration for use when a time lag of 30 sec for the indication of expired oxygen and carbon dioxide can be tolerated. The principal advantages of this instrument are a high degree of accuracy (error for oxygen, 0.05 vol%; for carbon dioxide, 0.007 vol%) and great stability in operation.

By supplementing the respiratory volume with ambient air (1, 16) so as to provide a constant flow from which an aliquot is fed to the analyzing devices for oxygen and carbon dioxide, the record can be calibrated directly for oxygen consumption and carbon dioxide discharge or, by applying Weir's nomogram (Fig. 1) for a given flow, in terms of calories per unit time.

REFERENCES

1. Benzinger, T.: Studies on respiration and gas exchange with continuous, direct recording methods, Ergebn. Physiol. 40: 1, 1938.
2. Benzinger, T., and Kitzinger, C.: A method for continuous recording of gas composition by means of an interferometer, Nav. Med. Res. Inst. Proj. no. NM 001 011: 1, 1948.
3. Berggren, G., and Christensen, E. H.: Heart rate and body temperature as indexes of metabolic rate during work, Arbeitsphysiol. 14: 255, 1950.
4. Blinn, K. A., and Noell, W. K.: The development of a method for continuous recording of alveolar carbon dioxide tension during hyperventilation test in routine EEG work, USAF School Aviat. Proj. 21-02-068, 1949.
5. Bothe, W.; Wollschitt, H., and Ruska, H.: Das zeiss'sche laboratoriums-interferometer als stoffwechselmessgerät, Arch. exper. Path. u. Pharmakol. 177: 635, 1935.
6. Clamans, H. G.: Continuous recording of oxygen carbon dioxide and other gases in sealed cabins, J. Aviation Med. 23: 330, 1952.
7. Consolazio, C. F.: Johnson, R. E., and Marek, E. (ed.): *Metabolic Methods* (St. Louis: C. V. Mosby Company, 1951).
8. Du Bois, A. B.; Fowler, A. C.; Soffer, A., and Fenn, W. O.: Alveolar CO_2 measured by expiration into the rapid infrared gas analyzer, J. Appl. Physiol. 4: 526, 1952.
9. Fowler, R. C.: A rapid infrared gas analyzer, Rev. Scient. Instruments 20: 175, 1949.

* American Meter Co., Inc., Albany, N. Y.

0. Fry, F. E. J.: A simple gas analyzer, Canad. J. Res. 27: 188, 1949.
1. Lilly, J. C.: Flowmeter for Recording Respiratory Flow of Human Subjects, in Comroe, J. H., Jr. (ed.): *Methods in Medical Research* (Chicago: Year Book Publishers, Inc., 1950), Vol. 2, pp. 113–121.
2. Luft, K.: Ueber eine neue methode der registrieren gas analyse mit hilfe der absorption ultraroter strahlen ohne spektrale zerlegung, Ztschr. f. techn. Phys. 24: 97, 1943.
3. Lusk, G. (ed.): *The Science of Nutrition* (4th ed.; Philadelphia: W. B. Saunders Company, 1928).
4. Rein, H.: Gas exchange recorder: Attempts at continuous registration of respiratory gas exchange in man and animals, Arch. f. exper. Path. u. Pharmakol. 171: 363, 1933.
5. Silverman, L., and Whittenberger, J. L.: Clinical Polyneumograph, in Comroe, J. H., Jr. (ed.): *Methods in Medical Research* (Chicago: Year Book Publishers, Inc., 1950), Vol. 2, pp. 104–112.
6. Spoor, H. J.: Application of the infrared analyzer to the study of human energy metabolism, J. App. Physiol. 1: 369, 1948.
7. Webb, P.: Air temperatures in respiratory tracts of resting subjects in cold, J. Appl. Physiol. 4: 378, 1952.
8. Webb, P.: The measurement of respiratory air temperature, Rev. Scient. Instruments 23: 232, 1952.
9. Weir, J. B. de V.: New methods for calculating metabolic rate with special reference to protein metabolism, J. Physiol. 109: 1, 1949.
0. Young, A. C.: CO_2 analyzer, 1952. USAF School of Aviation Medicine, Proj. no. 22-1301-0002.
1. Young, A. C.: A recording gas flowmeter, 1952. USAF School of Aviation Medicine, Proj. no. 22-1301-0002.
2. Young, A. C.; Carlson, L. D.; Quinton, W. F., and Burns, H. L.: Electronic polyneumograph, AF Tech. Rep. no. 6243, 1951.
3. Young, A. C. and Jones, W.: A rapid oxygen analyzer, unpublished.

ENERGY METABOLISM AND METABOLIC REFERENCE STANDARDS

A. T. MILLER, JR., *University of North Carolina*

EVERY METABOLIC PROCESS involves either the utilization or the liberation of energy. Total energy metabolism is the resultant of all reactions of both types and is commonly measured in terms of the associated heat production or oxygen consumption. Each of these indexes of metabolic rate has limitations which must be recognized if serious errors are to be avoided. For example, measurement of heat production is not a valid indication of metabolic rate when the body temperature is changing and oxygen consumption parallels metabolic rate only during a condition of steady state.

Two types of metabolic rates are commonly measured, the basal metabolic rate and the metabolic rate during activity. These are discussed separately since the techniques of measurement and the reference units employed usually differ. Methods of collecting, sampling and analyzing expired air have been adequately discussed in Volume 2 of this series (6). Partitional calorimetry likewise is not discussed, since it is of primary value in studies of heat exchange and temperature regulation rather than in the study of total energy metabolism. The question of the selection of suitable metabolic reference standards is discussed for 2 reasons: (*a*) the interpretation of metabolic data is often colored by the reference standard selected, and (*b*) recently developed techniques have made possible the more accurate measurement of various "metabolic body sizes."

I. Measurement of Energy Metabolism

A. BASAL METABOLIC RATE (BMR)

This is the energy expenditure of a subject lying quietly in bed after 8 hr sleep and at least 12 hr after the last meal. There is a definite practice effect, due probably to increasing muscular relaxation, so that the result is often too high on the first one or more tests. The magnitude of this training factor has been estimated at about 8% (19), so that it cannot be neglected in precise studies. In the usual procedure, little attention is paid to the diet before the test. However, Kleitman (12) found that the BMR of one of his subjects varied from 1,355 to 1,914 Cal/24 hr depending on the caloric and especially the protein content of the previous day's diet, presumably due to prolonged specific dynamic action.

Because of the expensive equipment required, direct calorimetry seldom used today. Instead, the method of indirect calorimetry, volving the measurement of basal oxygen consumption, is almost niversally employed. Since the calorific value of oxygen varies cording to the foodstuff undergoing oxidation, either the respira- ory quotient (RQ) must be determined from analysis of expired r and nitrogen excretion or an assumed value (about 0.85) must e used. The latter procedure may involve a maximum error of bout ±5% (10).

The closed circuit clinical method of determining BMR from the ope of a spirometer record requires no detailed comment, since mplete instructions accompany each instrument. Aside from in- rumental leaks (readily detected by the resulting concavity of he tracing), the chief source of technical error is in the interpreta- on of records in which the breathing is irregular. This difficulty ay be partially overcome by the following procedure (18): the bject is directed to make 3 maximal expirations during the course f the test. A line joining these 3 points approximates the true xygen consumption slope since in each case only the residual air emains in the lungs.

PEN CIRCUIT METHOD

The open circuit method involves collection and analysis of the xpired air. It is somewhat more accurate than the closed circuit ethod, since the RQ is determined rather than assumed, and the bjective error of estimating the slope of the oxygen consumption ne is eliminated.

a) The subject presents himself in the basal state. His age, eight without shoes and weight with minimal clothing are re- orded. He then rests in bed in a quiet place for at least 30 min. ulse rate and respiratory rate are counted several times during the tter part of the rest period, and the test should not be started un- l they have reached a plateau. Oral temperature is also recorded.

b) The subject is connected to the Tissot spirometer via a outhpiece or face mask and a 3-way stopcock so that he inspires om air and the expired air is collected for several minutes. The -way stopcock is then turned, disconnecting the subject from the irometer, and the spirometer bell is emptied by manual pressure. his procedure is repeated several times to wash out the dead space the spirometer.

c) The initial spirometer reading is recorded and at the end of a rmal inspiration the 3-way stopcock is turned so as to connect the bject to the spirometer and a timed collection of expired air is gun. This collection period should be at least 15 min long.

d) At the end of the collection period the 3-way stopcock is again turned to disconnect the subject from the spirometer, the spirometer reading, temperature and barometric pressure are recorded, and the mouthpiece or face mask is removed from the subject.

e) A sample of expired air is obtained from the spirometer and analyzed in duplicate for CO_2 and O_2.

Calculations.—a) Pulmonary ventilation. This is expressed as liters of air (at STP) expired/min and is calculated from the equation

$$PV = \frac{B - W}{760(1 + 0.0036t)} \times (SR_2 - SR_1)(SF)$$

where PV = pulmonary ventilation (liters/min), B = ambient barometric pressure, W = water vapor tension at spirometer temperature, t = spirometer temperature, SR_2 = final spirometer reading, SR_1 = initial spirometer reading and SF = spirometer calibration factor (liters of gas/cm of scale reading).

b) "True oxygen." This represents the number of milliliters of oxygen consumed for every 100 ml of air expired. When multiplied by the minute volume of ventilation, the oxygen consumption/min is obtained. When the RQ is 1.00, the volumes of inspired and expired air are the same since the oxygen consumed is replaced by an equal volume of carbon dioxide. When, as is usually the case, the RQ is less than 1.00, the volume of inspired air is greater than that of the expired air and must be calculated. Since the nitrogen content of inspired air is 79.04%, then

$$(1)\quad \text{Vol. insp. air} = \text{vol. exp. air} \times \frac{\%\ N_2 \text{ in exp. air}}{79.04}$$

$$(2)\quad \text{Vol. } O_2 \text{ insp.} = \text{vol. insp. air} \times \frac{20.93}{100}$$

$$(3)\quad O_2 \text{ consumed} = \text{vol. insp. air} \times \frac{20.93}{100} - \text{vol. exp. air} \times \frac{\%\ O_2 \text{ in exp. air}}{100}$$

$$= \text{vol. exp. air} \times \frac{\%\ N_2 \text{ in exp. air}}{79.04} \times \frac{20.93}{100} - \text{vol. exp. air} \times \frac{\%\ O_2 \text{ in exp. air}}{100} = \frac{\text{vol. exp. air}}{100} (\%\ N_2 \text{ in exp. air} \times 0.265 - \%\ O_2 \text{ in exp. air})$$

This last expression permits calculation of oxygen consumption from experimentally determined values of volume and composition of expired air.

c) Respiratory quotient

$$RQ = \frac{\%\ CO_2 \text{ in exp air} - 0.03}{\%\ N_2 \text{ in exp air} \times 0.265 - \%\ O_2 \text{ in exp air}}$$

nograms are available (7) for reading directly the values for ue oxygen" and RQ from an analysis of expired air.

d) Calorific value of oxygen. Protein metabolism is ordinarily glected, and the calorific value of oxygen is read from tables of nprotein RQ.

e) Basal metabolism. Surface area is read from standard tables ating surface area to height and weight. Metabolism, expressed Cal/sq m surface area/hr is compared with that of "normal" bjects of the same age and sex and the BMR is expressed as a us or minus percentage deviation from the normal.

B. METABOLISM DURING WORK

For work of moderate intensity, oxygen consumption may be de-mined by a modified closed circuit technique using large-bore bing and low resistance valves close to the face mask. The use-ness of the method is limited by the fact that a straight-line pe is obtained only during a steady state of oxygen consumption. For most experiments, work metabolism is more accurately easured by the open circuit method, with collection of expired air Tissot spirometers or in Douglas bags.

EN CIRCUIT METHOD

1. *Principle.*—Basal or resting metabolism is first measured as baseline for calculating the net energy cost of work. If the total ergy cost of work is to be determined, expired air is collected roughout the period of work and recovery, until oxygen consump-n approaches the baseline (usually about 60 min). The oxygen ot for 15 min recovery is roughly proportional to the total oxy-n debt and may be used without great sacrifice of accuracy.

2. *Apparatus.*—*a*) Face mask. Comfort, accurate fit and nimal dead space are essential. Suitable masks and sources of pply are listed in *Methods*, Volume 2, page 77.

b) Respiratory valves. These should have low opening pressure d minimal resistance to flow. The "J" valve* is satisfactory.

c) Stopcocks. These should be wide-bore, leakproof and easily rned. A 3-way aluminum stopcock† should be lightly greased th a lubricant such as Lubriseal and occasionally cleaned with ap and water.

d) Tubing. Long, narrow tubing, especially if corrugated, adds aterially to flow resistance. Suitable corrugated rubber tubing

* Warren E. Collins Co., 555 Huntingdon St., Boston.
† Arthur H. Thomas Co., Philadelphia; Cat. no. 6423.

(3.8 cm inside diameter) may be obtained in 75 cm lengths.‡ Minimal length consistent with maneuverability should be used.

e) Expired air collection reservoirs. The Tissot spirometer (capacity 600 liters) is the most satisfactory collecting system for stationary work experiments. Unfortunately many of these spirometers are equipped with small-bore (1.9 cm) tubing. This should be replaced by brass or stainless steel tubing of 3.8 cm inside diameter. An electric mixing fan should be installed at the top of the bell, and it is desirable that a baffle mixer of 10 liter capacity be installed at the inlet of the spirometer from which expired air samples may be drawn during exercise experiments. A complete description of the construction and use of this spirometer is given in *Methods*, Volume 2, page 94. For fractional collections of expired air, Douglas bags§ or Neoprene balloons‖ are satisfactory. Since these bags are not entirely impervious to carbon dioxide, expired air samples should not be stored in them longer than is necessary.

A portable, light-weight aliquot expired air sampler (K-M calorimeter) has been described by Kofranyi and Michaelis (13) which lends itself admirably to nonstationary work experiments when the size and bulk of the Douglas bag are objectionable.

Comment by Ray G. Daggs

The portable, light-weight aliquot expired air sampler (weighing about 10 lb) described by Kofranyi and Michaelis (13) is essentially a refinement of the old Geppert-Zuntz method using a dry gas meter strapped to the subject. An aliquot sampler built into the meter is in the form of a precision piston pump (travel volume of 3 cc) with a sliding valve which, when connected to the inlet (expired air from face mask) of the meter by a small tube, allows for a sampling of about 0.1% of the total expired air. The piston is operated by an eccentric directly connected to the diaphragm crank shaft of the meter. The outlet from the piston pump is a small tube to which a rubber bladder is connected. The loss of CO_2 by diffusion through the rubber bladder is prevented by placing it in an expired-air-filled can for transport to the laboratory for analysis. The bladder capacity is about 1 liter, which is adequate for about 1 hr of medium hard work. The apparatus does not produce a noticeable resistance to breathing and can be used on the actively working man since the calibration factor (measured against a wet gas meter) changes only from 1.04 for the stationary position to 1.09 under violent movement.

f) Expired air sampling vessels. (*See Methods*, Vol. 2, p. 122.) The selection of a sampling device depends partly on the method of gas analysis to be used. The Bailey bottle is perhaps best if the gas

‡ McKesson Appliance Co., Toledo, Ohio.
§ Arthur Thomas Co., Philadelphia.
‖ Dewey and Almy Chemical Co., Cambridge, Mass.

nple is to be stored before analysis. Oiled 50 ml syringes are ry convenient and are recommended if expired air is to be an- zed by means of the Pauling meter (*see* below).

g) Gas analysis apparatus. (*See Methods*, Vol. 2, pp. 125–137.) curate results are obtainable with the Haldane, Van Slyke or holander techniques. The chief disadvantage of these chemical thods is the time required when many analyses are to be per- med. If oxygen concentration alone is to be determined, the uling meter¶ is very rapid, though somewhat less accurate than methods described above. Behrmann (1) has described a tech- que for analyzing both carbon dioxide and oxygen with the uling meter. For following changes in expired air composition at quent intervals (30–60 sec), the thermal conductivity method of rg (2) is very useful.

3. *Procedure* (*using Tissot spirometer*).—*a*) With the subject a basal or resting state and in the posture to be used during the ercise period, expired air is collected in a Douglas bag for 10 min determination of baseline oxygen consumption.

b) The initial spirometer reading is recorded and at the signal begin exercise, the 3-way stopcock is turned, connecting the sub- t with the spirometer.

c) Spirometer readings are made every $^1/_2$ min and fractional pired air samples are obtained from the spirometer intake pipe ch minute (the sample must be obtained over a period greater an 20 sec) if the time-course of oxygen consumption is desired.

d) At the signal to end exercise, the 3-way stopcock is turned to nnect the subject to a Douglas bag and collection of expired air continued. As quickly as possible, the final spirometer reading d the temperature of the expired air in the spirometer are re- rded, a sample obtained for analysis and the spirometer emptied.

e) After a 3 min collection of expired air in the Douglas bag, the way stopcock is turned, connecting the subject to the spirometer ce more, and expired air collection is continued, preferably for otal recovery period of 60 min.

4. *Calculations*. *a*) Pulmonary ventilation, oxygen consump- n and RQ are calculated as described on page 76.

b) Oxygen debt is calculated by subtracting from the total oxy- n consumption during the recovery period the basal or resting ygen consumption for an equivalent period of time.

c) Total oxygen cost of exercise is the sum of the net oxygen nsumption during exercise and the oxygen debt.

¶ National Technical Laboratories, Pasadena 2, Calif.

II. Metabolic Reference Standards

The selection of suitable metabolic reference standards is of more than academic interest because it often colors the interpretation of experimental data. For example, if BMR is expressed in terms of surface area or of gross body weight, it is often subnormal in obese persons. When, however, the data are expressed in terms of fat-free body mass, the same subjects may be found to have an elevated BMR.

The almost universal use of surface area as the reference standard for basal metabolism is based partly on tradition, derived from a naive interpretation of the mechanism of heat loss from the body, and partly on the fact that in animals of widely varying size the metabolic rate based on surface area is more nearly constant than that based on body weight. Since it is no longer believed that heat loss from the body actually determines the rate of heat production, the "surface area law" is no longer tenable. Furthermore, there is good evidence that the anatomic surface area and the radiation surface area are not identical.

Two types of approach have been used in the attempt to devise a more logical unit for expressing basal metabolic rate. The mathematical approach derives an exponential power of the body weight from the regression equation relating basal metabolism to body weight. When various species of animals of widely different sizes are compared in this way, the best unit varies from $W^{0.70}$ to $W^{0.75}$ (11). When, however, different individuals of the same species are compared, the factor varies widely, e.g., from $W^{0.55}$ to $W^{0.80}$ for dogs (11, 9) and from $W^{0.44}$ to $W^{0.70}$ for women living in southern India (14, 8). It is almost inconceivable that all of these exponents measure a real metabolic body size.

In spite of this tremendous intraspecific variability, the body weight exponent is often assigned a precise physiologic significance. It cannot be emphasized too strongly that these exponents remain mathematical abstractions no matter how closely they may predict metabolic rates.

The second approach to a metabolic body size is based on the premise that the mass of actively metabolizing tissue is the logical reference unit, since the result then indicates the *intensity* of metabolism. Since most of the variability in body composition is due to fat, which is relatively inert metabolically, the lean body mass (body weight minus fat) is the simplest derived unit. Lean body mass may be calculated from experimentally determined values of body specific gravity (17), total body water (17) or urinary creatinine excretion (16). In a study on 48 college students

5), lean body mass was found to predict basal oxygen consumption more closely than either surface area or gross body weight. n even more rigorous metabolic unit is the Minnesota "active ssue mass" (10), consisting of body weight minus the sum of body t, extracellular fluid and bone minerals. Data for comparison of an body mass and active tissue mass are not available.

ETERMINATION OF LEAN BODY MASS

1. *From total body water.*—*a*) Principle. The lean body mass is haracterized by a relatively constant water content, about 72% 17), whereas fat contains very little water; therefore the greater he fat content of the body, the lower the body water content. his concept has been validated by direct chemical analysis of ody fat.

b) Procedure. This is adequately described elsewhere in this eries (4).

2. *From body specific gravity.*—*a*) Principle. The specific ravity of fat is lower than that of any other body constituent. Hence the higher the body fat content, the lower the body specific ravity. The body specific gravity is calculated from the weight of he body (1) in air and (2) submerged in water, by application of Archimedes' principle.

b) Apparatus. The *water tank* is constructed of 1/4 in. sheet netal, or similar material, painted inside and outside with waterroof paint, and is approximately 5 ft in diameter and 5 ft in epth. It is provided with a floor drain for emptying. A satisfactory *cale* is the model 1851-Y Yale Hanging Type Scale,* having a dial apacity of 20 kg graduated by 20 g. The *seat* is made of stainless teel and is suspended from the scale hook. The scale and seat are aised and lowered by means of a Whiting electric *hoist*,† operated y a 1 horsepower, 220 v, 3-phase electric motor.

c) Procedure. The subject is seated and, with the breath held n maximal expiration, is submerged until only the scale hook reains above water level. The scale is unlocked, the submerged eight recorded, the scale locked again and the subject brought to he surface by means of the hoist. The entire procedure requires bout 10–15 sec. It is essential that the subject remain motionless hile the weight is being read to minimize oscillations of the scale ndicator. The procedure should be repeated until several successive weighings agree within 0.1 kg. The first several weights recorded are frequently low because of incomplete expiration and the

* Yale and Towne Manufacturing Co., Philadelphia.

† Atlas Equipment Co., 229 Southwest Blvd., Kansas City, Mo.; model 451H30, apacity 500 lb, hoisting speed 30 ft/min.

resulting increased buoyancy of the lung air. Residual air is determined by one of the standard methods described in Volume 2 of this series (6). Residual air measured while the subject is submerged is slightly less than when not submerged (5), but the error introduced by using nonsubmerged residual air values is not serious. When large numbers of subjects are to be studied, the average residual air volumes for the age and sex of the subjects (6) may be used instead of experimentally determined values with an average error not exceeding 1 or 2% in the final body fat content.

d) Calculation.

$$\text{Body sp. gr.} = \frac{\text{wt. in air (kg)}}{\text{wt. in air} - (\text{wt. submerged} + \text{resid. air in liters})}$$

$$\text{Body fat content (\%)} = 100\left(\frac{5.548}{\text{sp. gr.}} - 5.044\right)$$

(Body fat content is more conveniently read from a table constructed from a large-scale graph of the equation.)

3. *From urinary creatinine excretion.—a*) Principle. Even though the precise significance of urinary creatinine is uncertain, it appears to be related to the resting metabolism of muscle, which forms a large fraction of the lean body mass. In a recent study (16) it was found that values of lean body mass predicted from standardized creatinine excretion and calculated from body specific gravity measurements agreed, in 90% of cases, within ±13.1%. For groups of 50 or more subjects the group values by the 2 methods should agree within ±2.0%.

b) Procedure. Accurately timed urine collections are made over a 4 hr period (preferably from 8:00 A.M. to noon) on each of 5 successive days. Urine creatinine is determined by Folin's alkaline picrate method. Lean body mass (LBM) is calculated from the equation

$$\text{LBM (in kg)} = 20.97 + 0.5161 \text{ creatinine (mg/hr)}$$

Comment: The 4 hr collection, suggested by Best *et al.* (3), gives more constant results than does a 24 hr collection for reasons which are not clear. Although daily variation in creatinine excretion may vary as much as 20% in the same individual, the average of 5 successive collections is reasonably constant from 1 week to another.

The *energy cost of activity* is more closely related to body weight than it is to surface area when the activity involves moving or lifting the body. The *intensity of activity* is best expressed as the ratio of the energy requirement of activity to the basal or resting energy requirement. The *energy cost of performing external mechanical work* is expressed in terms of Cal/kg-m of work performed.

mment by Grover C. Pitts

The most provocative portion of Dr. Miller's paper is that concerning etabolic reference standards. As he has indicated, basal heat production $W^{0.70-0.75}$ is almost independent of total body size in comparing a ide variety of mammalian species. By contrast, studies on a given ecies have generally yielded a distressing range of "standards." This tuation appears paradoxical since in the latter instance species differences have been eliminated. However, the interspecific studies enerally deal with mean values, a procedure which statistically minizes individual variations in body fat. On the other hand, in intraecific studies individual variations in body fat contribute heavily to ie over-all variability.

An unpublished series of determinations on guinea-pigs reveals a positive correlation between body weight and per cent body fat, a circumtance which, it is reasonable to believe, will obtain also in man. Thus, ith increasing total body weight there is an increase in the percentage f material which does not contribute significantly to energy metabolism, s well as an increase in the mass of actively metabolizing tissue. Hence ie variation of body fat in a given species, which Dr. Miller emphasizes, ot only contributes to the over-all variability of a series of measureents but can actually distort the curve relating metabolic rate to body eight in a misleading way.

REFERENCES

1. Behrmann, V. G., and Hartman, F. W.: Rapid CO_2 determination with the Pauling O_2 analyzer, Federation Proc. 10: 12, 1951.
2. Berg, W. E.: Individual differences in respiratory gas exchange, Am. J. Physiol. 149: 597, 1947.
3. Best, W. R.; Kuhl, W. J., Jr., and Friedemann, T. E.: Diurnal trend and variation of urinary creatinine excretion, Federation Proc. 11: 188, 1952.
4. Brodie, B. B.: Measurement of Total Body Water, in Visscher, M. B. (ed.): *Methods in Medical Research* (Chicago: Year Book Publishers, Inc., 1951), Vol. 4.
5. Brozek, J.: Henschel, A., and Keys, A.: Effect of submersion in water on the volume of residual air in man, J. Appl. Physiol. 2: 240, 1949.
6. Comroe, J. H., Jr. (ed.): Pulmonary Function Tests, *Methods in Medical Research* (Chicago: Year Book Publishers, Inc., 1950), Vol. 2.
7. Consolazio, C. F.; Johnson, R. E., and Marek, E. (ed.): *Metabolic Methods* (St. Louis: C. V. Mosby Company, 1951).
8. Cullumbine, H.: Heat production and energy requirements of tropical people, J. Appl. Physiol. 2: 640, 1950.
9. Galvao, P. E.: Heat production in relation to body weight and body surface. Inapplicability of the surface law on dogs of the tropical zone, Am. J. Physiol. 148: 478, 1947.
10. Keys, A., *et al.*: *The Biology of Human Starvation* (Minneapolis: University of Minnesota Press, 1950).
11. Kleiber, M.: Body size and metabolic rate, Physiol. Rev. 27: 511, 1947.
12. Kleitman, N.: Basal metabolism in prolonged fasting in man, Am. J. Physiol. 77: 233, 1926.

13. Kofranyi, E., and Michaelis, H. F.: Ein tragbarer Apparat zur Bestimmung des Gasstoffwechsels, Arbeitsphysiol. 11: 148, 1941.
14. Mason, E. D., and Benedict, F. G.: Basal metabolism of South Indian women, Indian J. M. Res. 19: 75, 1931.
15. Miller, A. T., Jr., and Blyth, C. S.: Lean body mass as a metabolic reference standard, J. Appl. Physiol. 5: 311, 1953.
16. Miller, A. T., Jr., and Blyth, C. S.: Estimation of lean body mass and body fat from basal oxygen consumption and creatinine excretion, J. Appl. Physiol. 5: 73, 1952.
17. Osserman, E. F.; Pitts, G. C.; Welham, W. C., and Behnke, A. R.: In vivo measurement of body fat and body water in a group of normal men, J. Appl. Physiol. 2: 633, 1950.
18. Ryder, H. W., and Esselborn, V. M.: The determination of basal metabolism by periodic maximal exhalations, J. Lab. & Clin. Med. 34: 1742, 1949.
19. Vogelius, H.: Basal metabolism of girls and the use of metabolic standards, Acta med. scandinav., Supp. 165, p. 1, 1945.

RADIOMETRIC METHODS FOR MEASUREMENT OF SKIN TEMPERATURE

JAMES D. HARDY *and* ALICE M. STOLL, *Cornell University*

THE IMPORTANCE of skin temperature in the determination of heat exchanges between man and his environment has long been recognized. For significant results in the study of heat loss from the body, the average skin temperature over the entire body surface must be measured. Accuracy and rapidity of measurement of skin temperature are vital not only to studies of the physiologic responses to heat and cold but also to investigations of skin circulation and of thermal skin damage and pain.

Many devices and methods have been developed for the measurement of skin temperature and have yielded equivocal results because of the technical difficulties involved. Foremost among these difficulties is the proper calibration of an instrument. Since the temperature to be measured is that of a boundary between 2 different mediums the difficulty of calibration of surface temperature-measuring devices, under actual conditions of use, is readily appreciated. The problem is further complicated by the variable effects of the skin thermometer itself in changing the surface temperature as it makes a measurement. The thermal capacity of the applicator, the pressure with which it is applied, the sweat on the skin surface, are but a few of the factors to be considered. Radiometric devices, which do not come into contact with the skin surface, largely avoid these dangers and have been developed to provide rapid (about 0.2 sec), accurate measurements of exposed skin surfaces. Radiometric instruments are not yet adapted to measurements under clothing.

Table 1 gives a comparison of several skin temperature thermometers when tested against a leather surface of known temperature. Test conditions were arranged to simulate conditions met with in the laboratory. When the thermocouples and other types of contact instruments are applied to the skin surface rather than to a leather surface, the variability in their readings becomes greater than that shown in Table 1. The decrease in dependability has been shown to arise from the effects of pressure of the thermometer on the skin. Even the most ingenious devices for the regulation of contact pressure do not afford appreciably better measurements of skin temperature. For these reasons, the emphasis of this discussion

TABLE 1.—COMPARISON OF SEVERAL SKIN TEMPERATURE THERMOMETERS ON LEATHER SURFACE OF KNOWN TEMPERATURE

AVERAGE DEVIATION FROM EXTRAPOLATED TEMPERATURE - °C.

INSTRUMENT	EXPERIMENTAL CONDITIONS				
	ROOM (NORMAL)	WIND VELOCITY 4 FT./SEC.	INFRA-RED RADIATION (HOT STOVE)	1500 WATT LAMP RADIATION	1500 WATT LAMP WIND VELOCITY 2 FT./SEC.
"DERMAL" RADIOMETER	−0.04	+0.02	−0.01	−0.04	−0.05
THERMOCOUPLE #40 GAGE WIRE (BARE)	+0.05	−0.28	−0.10	−0.04	−0.12
THERMOCOUPLE #28 GAGE WIRE (BARE)	+0.13	−0.27	−0.49	−0.49	−0.40
THERMOCOUPLE SOLDER BEAD (ADHESIVE TAPE)	+0.14	−0.08	−0.18	−0.91	−0.58
THERMOCOUPLE #40 GAGE WIRE (GLUED)	+0.22	−0.26	+0.03	−0.37	−0.42
"DERMALOR" RESISTANCE THERMOMETER	−0.44	+0.50	−1.75	−1.18	+0.11
"PYROMETER" STRIP THERMOCOUPLE	−3.2	−3.0	−5.9	−4.9	−2.7
"PYROMETER" SOLDER BEAD	−5.3	−3.7	−7.2	−6.8	−3.7
COPPER MESH THERMOCOUPLE	−0.36	−0.57	−0.42		
DISC THERMISTOR	−1.21	−1.22	−1.91		
RUBICON SKIN THERMOMETER	−2.85	−0.84	−4.93		

is placed on the recent developments in radiometric methods for measuring skin temperature.

I. Rapid Measurement of Skin Temperature during Exposure to Thermal Radiation

As already noted, reliable measurements of skin temperature are difficult to obtain even in the "steady state," and no method has been available for making such measurements when there are rapid fluctuations in skin temperature. Recently Henriques (5) and Buettner (1) have developed theoretical equations relating skin temperature change to intensity and duration of exposure to thermal radiation. This has given new impetus to experimental study of skin temperature during the course of such exposures, and we describe

here a method employing well tested principles for rapid measurement of skin temperature during exposure to intense visible and near infra-red radiation.

The radiometric technique of measuring skin temperature involves the comparison of the radiation from the skin with that from a black body of known temperature. However, measurement was slow due to long response times of the radiometer and the galvanometer. Recent developments in detection of infra-red radiation have included the Golay pneumatic detector (2) which maintains a high level of sensitivity even when used with exposures as short

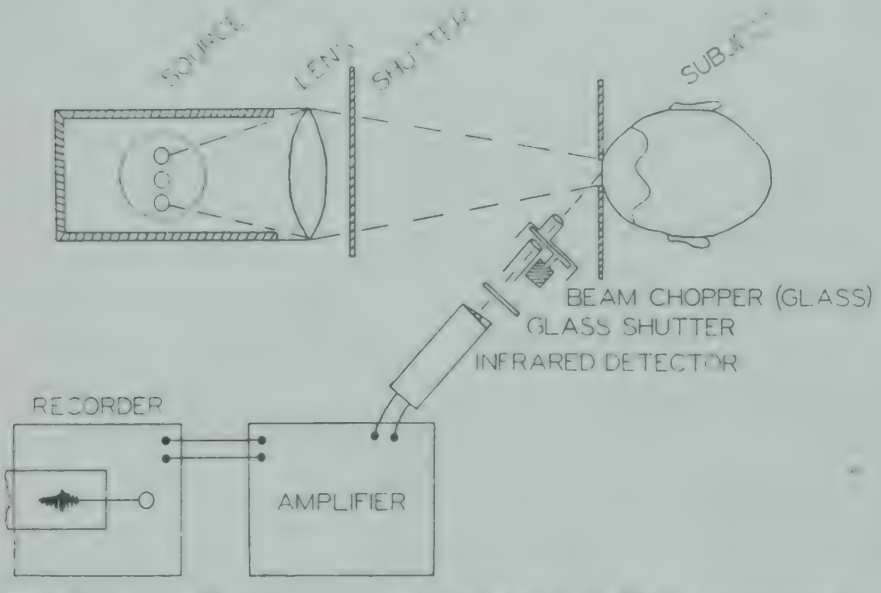


FIG. 1.—Experimental arrangement for measuring skin temperature during exposure to thermal radiation.

as 0.01 sec. The use of this detector seemed to offer a solution to the problem of rapid skin temperature measurement.

PROCEDURE

The arrangement of apparatus for exposure of the skin to known amounts of radiation and for measuring the changes in surface temperature before, during and after such exposure is shown in Figure 1. The source of radiant heat was a 500 w projection lamp mounted in a ventilated housing. The light from this lamp was focused by condensing lenses upon the forehead of the subject so as to illuminate uniformly an area of skin 4 × 4 cm. An opaque shutter, operated by hand, was mounted between the heat source and the subject. Intensity of radiation could be controlled by altering the electrical input to the lamp and was measured by placing a calibrated radiometer within the aperture in place of the subject's forehead.

The device for measuring skin temperature consisted of the Golay infra-red detector which was connected through a suitable amplifier to an oscillographic recorder. Between the detector and the forehead a series of diaphragms was arranged so that the detector "saw" only a circular area of skin 7 mm in diameter in the exact center of the irradiated area. The area of skin studied was limited in this way so as to avoid the effects of lateral conduction of heat in skin at the edges of the aperture (4). In the path of the beam of radiation from the skin to the detector were mounted a glass shutter and a glass interrupter or "chopper." The chopper was a semicircular disk of clear glass which interrupted the beam 5 times/sec. The

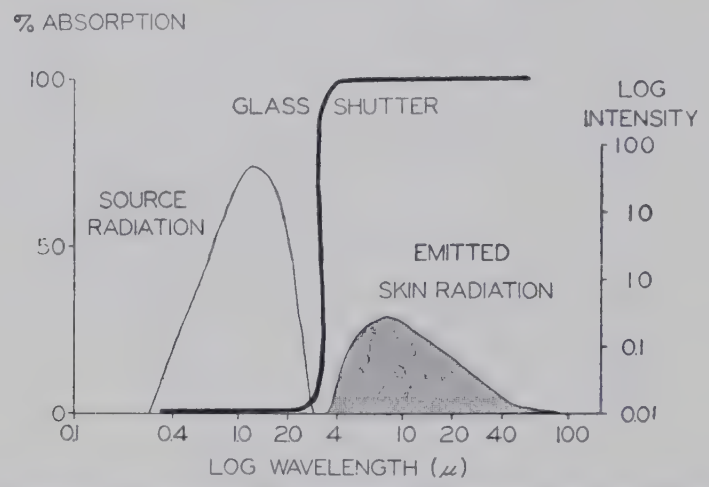


FIG. 2.—Diagram of spectral distribution of energy emitted from radiation source and from skin. Ordinate on right is log intensity in millicalories (0.001 g-cal)/sec/sq cm. Absorption curve of glass is shown to indicate how the source radiation passes through the glass disk interrupter and is thus not "chopped," whereas the emitted radiation from the skin is not transmitted and is therefore "chopped."

glass chopper was used rather than the usual metal one because it was necessary to reduce the effect of the reflection of light from the skin when the blackened skin was irradiated. Figure 2 shows how this was accomplished.

The spectral energy distribution of radiation from the radiating source was measured as shown in Figure 2 and is labeled at the left in Figure 3. This radiation passed through the glass disk without being appreciably absorbed so that as the disk rotated there was no alternating signal to be amplified and recorded. On the other hand, the radiation which was emitted from the skin and which must be measured to ascertain the skin temperature was not transmitted by the glass disk and therefore was periodically occluded as the disk rotated. This furnished an alternating voltage

from the detector, which could be amplified and recorded. The advantage to be gained from such an arrangement is demonstrated in Figure 4. To the left the source radiation reflected from the skin is shown as interrupted by a metal disk and by a glass disk. The reduction of effect of this radiation on the detector by using the glass disk was roughly 7:1; the residual effect was due to the reflection of light from the glass disk. The longer wavelength radiation emitted by the skin was interrupted as well by the glass disk as by the metal disk and thus with the glass disk its effect on the detector was magnified 7-fold relative to the diffusely reflected radia-

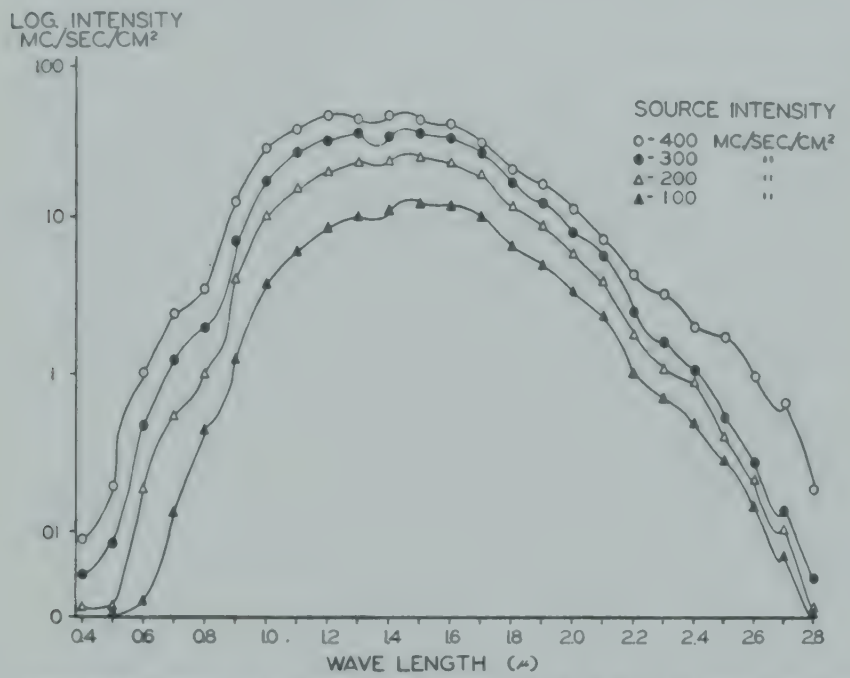


FIG. 3.—Spectral distribution of the energy (0.41–2.8 μ) from the source of radiant heat at various intensities.

tion. An even further magnification can be obtained by treating the surface of the glass disk so as to be nonreflecting. (This has not been necessary to prosecute investigations involving the low reflecting power of blackened skin.)

Theory.—The operation of this radiometer requires that the temperature of the surface of the glass shutter (facing the detector) be known. A housing was built around the interrupter and a thermocouple mounted near this surface of the disk as it rotates. The detector then compares the radiation from the skin with that from the inner surface of the disk 5 times each second.

The glass shutter shown between the interrupter and the detector permits measurement of the residual reflection from the source

which is transmitted through this shutter. The radiation emitted from the skin is excluded when the glass shutter is in the light pathway.

During irradiation of the skin, and with the glass shutter removed, the detector will receive both the energy emitted by the skin and that reflected from the skin surface. This can be expressed in terms of the deflection produced on the recorder as:

$$D_1 = K(S_0\epsilon_1\epsilon_2(T_1^4 - T_2^4) + \alpha I_0) \quad (1)$$

in which D_1 = deflection recorded with glass shutter removed; K = proportionality constant dependent mainly on the sensitivity of the detector; S_0 = Stefan-Boltzmann constant; T_1 = skin temperature (degrees Absolute); ϵ_1 = emissivity of India ink-blackened

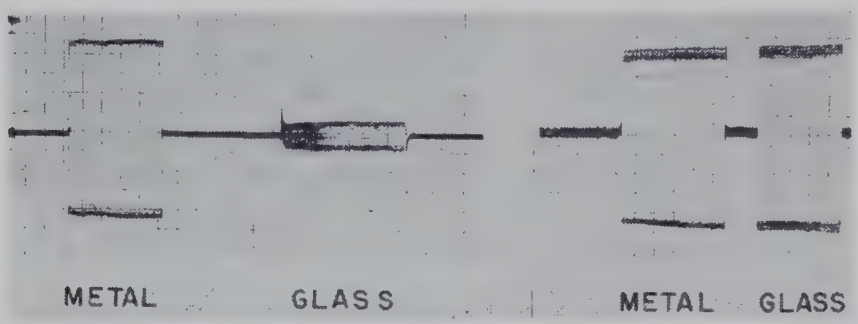


FIG. 4.—Effects of metal and glass choppers. *Left*, reflected source radiation: comparative recordings show minimizing of effect of reflected radiation by use of glass rather than the customary metal interrupter. *Right*, emitted skin radiation: recordings show no loss of sensitivity as regards emitted radiation from skin when metal interrupter is replaced by glass.

skin; T_2 = temperature of glass disk interrupter (degrees Absolute); ϵ_2 = emissivity of glass interrupter (chopper); α = reflection constant for blackened skin; I_0 = intensity of energy irradiating skin.

With the glass shutter in front of the detector, the radiation from the disk interrupter and skin is excluded and the reflected beam passes (with loss of 8% due to reflection by the shutter) to the detector. Thus the deflection in this instance will be

$$D_2 = K(\alpha I_0)0.92 \quad (2)$$

The skin temperature can be determined from the 2 equations to a sufficient degree of precision as

$$T_2 = T_1 + \frac{D_1 - 1.1D_2}{K'} \quad (3)$$

where $K' = 4S_0\epsilon_2\epsilon_1T_1^3K$. When there is no radiation falling on the skin, the temperature is given simply by

$$T_2 = T_1 + \frac{D_1}{K'} \tag{4}$$

The value of ϵ_1 deserves some mention. The effective emissivity of the normal human skin has a value of 0.985 (3), which is usually taken as being unity, i.e., black body radiator. However, when the skin is coated with India ink it becomes less black. Correction must be made for the deviation from black body radiation. This is done

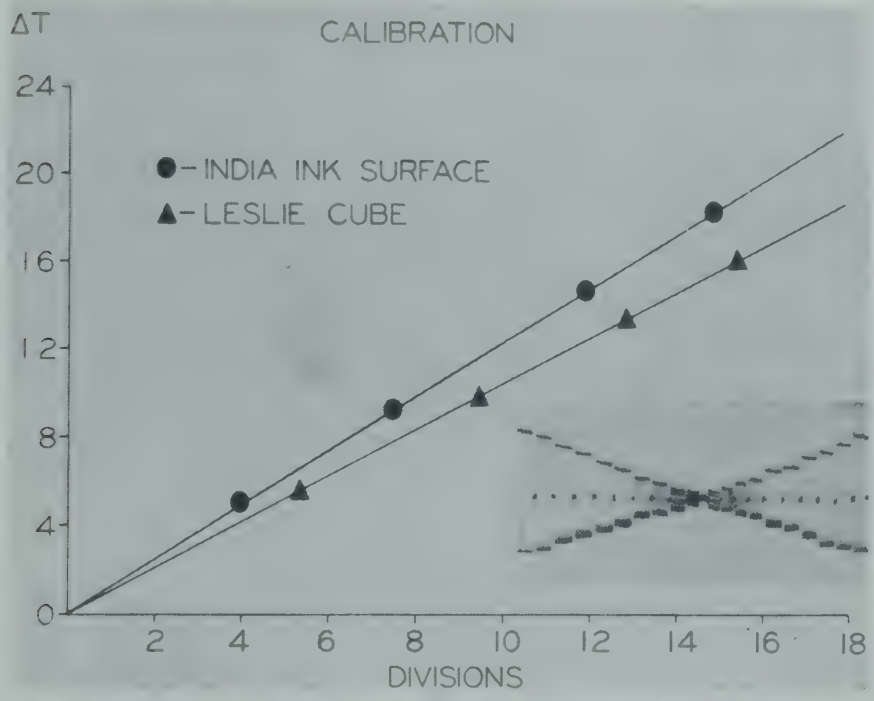


FIG. 5.—Typical calibration curve. ΔT = temperature of cube or inked surface minus temperature of glass disk. Division = amplitude of recorder oscillation. At lower right is a typical calibration record.

most simply by calibrating the radiometer with a surface coated with India ink.

Standardization.– Calibration of the instrument is carried out with a Leslie cube. The cube, filled with water at 50 C, is positioned in the aperture in place of the subject's head. The glass shutter is raised and the amplitude of the alternating voltage recorded by the oscillograph. The temperatures of the disk interrupter and the cube are read. The cube is allowed to cool a few degrees and the procedure is repeated until the cube and disk are at approximately the same temperature. Typical calibration curves are shown in Figure 5 and the actual recording made during calibration with the Leslie cube is shown at the lower right. Two curves are shown, 1 with the black body radiation from the cone of the cube and the

other with the copper side of the cube roughened with emery paper to simulate the skin, and thoroughly coated with India ink. The importance of the emissivity difference is thus demonstrated. For experiments in which the surface studied was coated with India ink, the calibration used was that given by the upper curve in Figure 5. With unblackened skin the lower calibration curve was used.

The reproducibility of the calibration was ± 0.1° C. In practice, a warm-up period of 30–60 min should be allowed before a calibra-

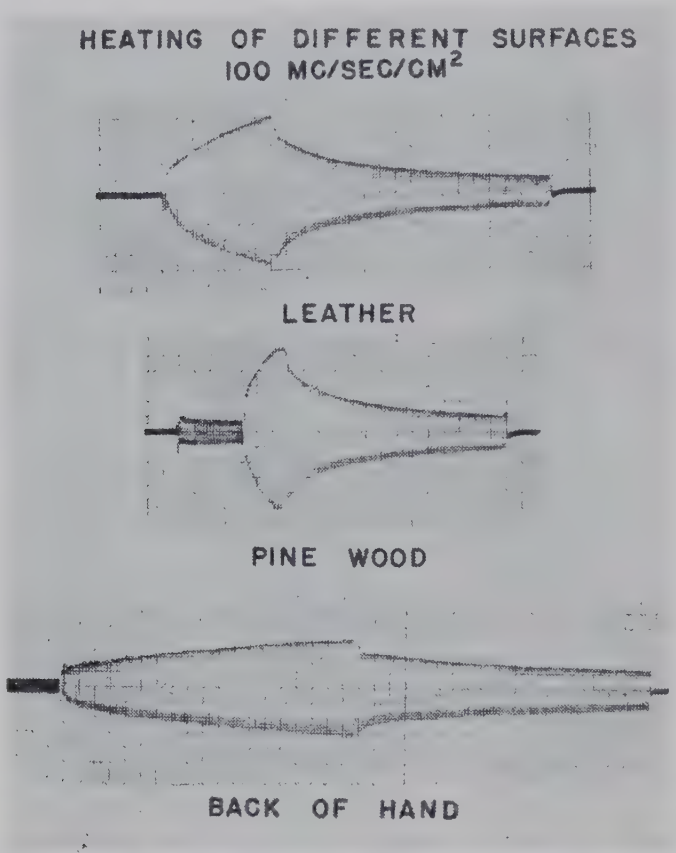


FIG. 6.—Recordings of temperature elevation of different surfaces before, during (period of increasing amplitude) and after exposure to thermal radiation. Time marked in seconds.

tion is begun. During the course of several hours, battery drain in the amplifying circuits sometimes caused changes in the calibration curve, and for this reason repeated calibrations were done during long experiments. A calibration usually required about 10 min to complete.

Response time.—As mentioned earlier, the radiometer compared the radiation from the skin and disk 5 times/sec. The complete response time of the amplifier was 0.05 sec, that of the detector 0.01 sec, and that of the recorder 0.02 sec. Thus, interrupting the

beam of radiation at 0.2 sec intervals is slow enough to allow complete equilibrium to be attained for each deflection. With some loss of sensitivity, an interruption rate of 100/sec is possible. In the present arrangement the sudden opening of the opaque shutter to the radiant heat source always caused an original deflection which had to be discarded (see Fig. 6). However, by the second or third oscillation the recorder reading was dependable. Thus, as used in the experiments, the effective complete response time was judged to be 0.4–0.6 sec.

Typical results.—Records of change in surface temperature during irradiation are shown in Figure 6. All surfaces were coated with India ink before exposure so that the data would be comparable as regards reflecting power for the visible and near infra-red and as regards emissivity in the far infra-red. The recordings were made in the following way. The surface to be measured was put into place in the aperture after allowing the ink to dry. The glass shutter was then raised and the initial temperature of the surface recorded. This is shown by the oscillations at the left of the records. After a few seconds (each square = 1 sec) the opaque shutter (*see* Fig. 1) was lifted and the surface irradiated at the rate of 100 mc/sec/cm^2 for as long as the deflection could be recorded, or, as in the lowest record, until the subject reported pain. The opaque shutter was then inserted, cutting off the radiation, and cooling of the surface followed. Finally, the glass shutter was lowered, excluding the emitted radiation from the surface. The rapid heating of the pine wood and leather in comparison to the skin is evidence of the different thermal diffusivities of the substances irradiated.

II. Portable Radiometer for Rapid Measurement of Skin Temperature over Small Areas*

The utilization of thermistors has permitted the elimination of the relatively slow galvanometer-potentiometer system in radiometric devices and made possible a portable, rapid-response radiometer suitable for the measurement of skin temperatures of areas such as the finger-tip, individual toes, footpads of small animals, etc.

APPARATUS

Principle.—The simplest of these instruments consists of a resistance bridge formed of 2 high-resistance flake thermistors and 2 fixed resistances. It is constructed so that 1 thermistor is exposed

* Dermal Radiometers can be obtained from the Co-Design Corporation, 751 Main St., Winchester, Mass.

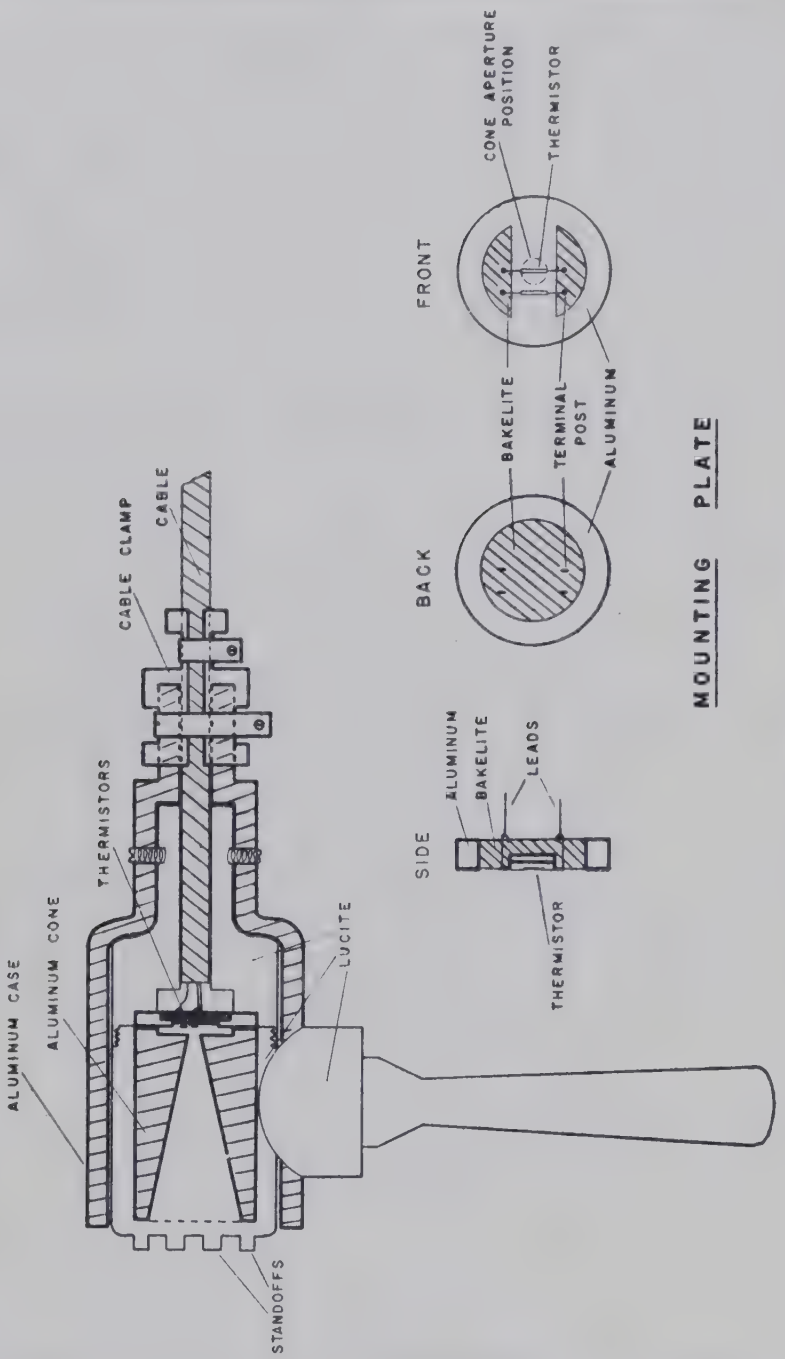


FIG. 7.—Radiometer assembly.

to radiation from the skin while the other is hidden from it and both thermistors are mounted in the same plane. Thus, they are equally affected by conduction and convection and only the radiation affects the bridge balance. The output from the bridge is electronically amplified and fed into a microammeter calibrated to indicate the skin temperature directly in degrees.

A measurement is made in the usual manner. The radiometer is exposed to a reference black body radiator of known temperature

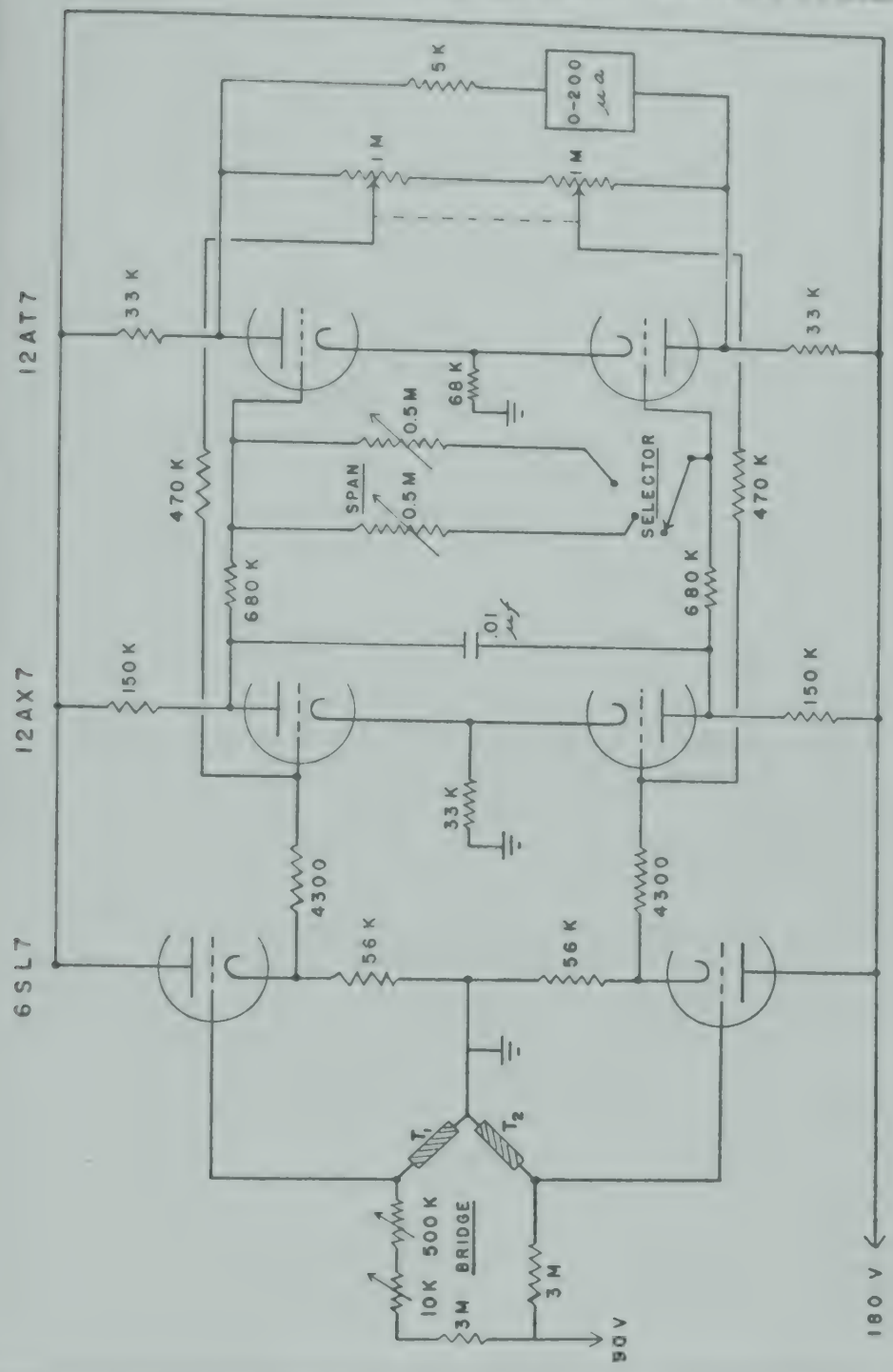


FIG. 8.—Radiometer circuit diagram.

and the meter needle is set at this temperature on the scale. The radiometer is then exposed to the skin and the temperature is indicated directly on deflection of the needle.

2. *Construction.*—A circular metal holder 2 cm in diameter and 3 mm thick is hollowed out to a depth of 1 mm through a radius of 9 mm. Two rectangular sections on opposite sides of the

inner circle are drilled through the holder. Bakelite wedges are forced into these spaces and extend 0.5 mm into the depression. Two holes are then drilled through each of these Bakelite sections and #24 B & S gauge copper wire is forced through the holes to form terminal posts which are held firmly in place with Ambroid cement to prevent rotation or slipping. The wire protruding on the "front side" of the Bakelite blocks, i.e., the side on which the thermistors are to be mounted, are trimmed to about 0.5 mm above the block, bent over and flattened slightly. Into the back of each Bakelite block and a little to 1 side of the wires, 2 screws are threaded and serve as connection posts for the leads from the thermistors and to the amplifier. The thermistors, 2.5 mm long, 0.01 mm thick and 0.2 mm wide, are now mounted side by side across the depression in the holder by suspension on silk fiber supports glued to the rim of the holder. The leads from the thermistors are soldered to the copper posts.

A polished aluminum cylinder 2 cm in outside diameter, having a cone 1.5 cm in diameter at the base and an opening 3 mm in diameter at the apex, is mounted in front of the thermistors so that 1 thermistor lies at the apex of the cone and the other is hidden under the rim of the cylinder. The cone and the mounted thermistors are held in place by a 2-piece Lucite cylinder which fits over the aluminum cylinder and the thermistor holder. The Lucite sections are threaded so that they can be screwed together and tightened down on the cone and mount to hold them firmly in place. An aperture in the rear section of the Lucite permits the leads to be brought out from the holder to the amplifier. The leads to the amplifier are thoroughly shielded and the common lead from the thermistors is soldered to this shield. The entire unit is mounted in an aluminum case having a wooden handle so that heat from the hand does not affect the sensitive elements when a reading is being made (Fig. 7). Over-all dimensions are 8.0 cm long and 3.5 cm in diameter. The aperture exposed to the test surface is 1.4 cm in diameter; thus, an area of 1.54 cm^2 is "seen" by the exposed thermistor. This is about $^1/_6$ of the area "seen" by the Dermal Radiometer.

3. *Characteristics.*—The instrument is well compensated for ambient temperature changes and shows no zero drift during use. The response time exclusive of the inertia of the microammeter is approximately 20 msec. Indications on the meter are for all practical purposes instantaneous (6). The amplifier circuit is shown in Figure 8. The sensitivity is such that full scale deflection on a 0–200 μa meter is obtained on exposure of the radiometer to a test surface 8° C different from that of the reference body. This is

more than enough for the purposes stated. The accuracy of the measurement is limited by the accuracy with which the scale may be read: for a 20° C span this is ±0.1° C. Greater accuracy may be obtained by increasing the scale length or reducing the span.

4. *Calibration.*—The radiometer is calibrated in the following manner. Two Leslie cubes (black body radiators) are prepared, 1 to maintain a temperature at or very close to the room temperature, and exactly at the temperature of the lower limit of the smallest temperature span desired; for skin temperature this is 20-40° C. The other cube is maintained at the exact temperature of the upper limit of the desired span. The temperature of each cube is measured by means of a constantan-copper thermocouple. The radiometer is exposed to the cube of lower temperature and the bridge balanced by means of the zero adjust potentiometer so that the meter reads zero. The radiometer is then exposed to the cube of higher temperature and the output adjusted by means of the calibrating potentiometer so that the meter reads full scale. This procedure is repeated until exposure of the radiometer to the cold cube yields a reading of exactly zero and exposure to the warm cube, a reading of exactly full scale without adjustment of the calibrating potentiometer. This initial calibration yields the necessary data for construction of the basic scale.

The output of the instrument is linear over short spans of temperature (20-30° C); however, to read the meter in degrees directly it is necessary to convert the linear scale to a fourth-power difference scale covering the temperature range desired. The scale intervals are laid off according to the following relationship

$$\sphericalangle i = \frac{\sphericalangle t}{(T_F^4 - T_0^4)} \times (T_1^4 - T_0^4)$$

where $\sphericalangle i$ = interval angle (in decimals); $\sphericalangle t$ = total angle described by the needle in moving from zero to full-scale deflection; T_F = highest temperature to be indicated (°K); T_0 = lowest temperature to be indicated (°K); T_1 = any intermediate temperature to be indicated (°K).

The instrument may be calibrated for 1 or more temperature spans in this manner.

5. *Discussion.*—This thermistor radiometer is entirely satisfactory for the measurement of radiant temperatures in ambient temperatures down to about 10 C. However, at lower ambient temperatures the inequality in the resistance in the thermistor arms with respect to the fixed resistance arms becomes sufficiently large to reduce the sensitivity of the instrument. To overcome this difficulty, another model has been constructed in which the bridge is

composed of 4 thermistors; thus the resistances of the bridge elements are equal at all ambient temperatures. This instrument also has much greater sensitivity, but for ordinary purposes such as clinical skin temperature measurements and surface temperature measurements under normal laboratory conditions the 2-thermistor instrument is preferable since it is less expensive, easier to construct and more rugged.

Comment by George W. Molnar

The radiometric techniques here described are ingenious and necessary in special cases when skin surface temperature should be measured with great accuracy or when application of a measuring element to the skin surface could change the properties of the surface. These special cases would include studies of thermal pain, burns, and cooling due to sweating on small areas. In almost all other studies, particularly general environmental studies, however, the advantages of a radiometer are often offset by the disadvantages. As the authors mention, there is no suitable radiometer for use under clothing, yet in many practical instances it is necessary to have the skin surface covered.

Air movement at the skin surface is usually not known because it is not readily measurable; yet the flow of air, and therefore convective heat exchange, at the skin surface can be markedly changed by a small change in position of the body. The environmental conditions, therefore, may not be defined with precision commensurate with skin surface temperature measurements of high accuracy.

Blood flow to the skin (differentiated from the total mass of the segment under consideration) is not necessarily measured by the conventional volume plethysmograph. Except when the skin surface temperature is rising, one cannot be certain that the blood flow even to a warm skin is sufficient to transport heat to the skin faster than heat moves by simple conduction alone; therefore it may be necessary to make cooling measurements following occlusion of arterial flow, and temperature measurements with high accuracy are not necessary for these tests.

In many cases in environmental studies it is important to have frequent, if not continuous, measurements at many places on the skin surface, and this can be done more conveniently with thermocouples and a recording potentiometer than with a radiometer. Finally, there is little point in making extremely accurate measurements when, in using them, one introduces errors of unknown magnitude such as in the calculation of so-called average skin temperature or of heat exchange from the skin. A simple thermocouple and galvanometer or potentiometer for measuring skin temperature can give all that is usually needed in most environmental studies. For specialized cases, as already mentioned, the radiometric techniques described by the authors are particularly suitable because of their great accuracy.

REFERENCES

1. Buettner, K.: Effects of extreme heat and cold on human skin: I. Analysis of temperature changes caused by different kinds of heat application, J. Appl. Physiol. 3: 691, 1951.
2. Golay, M. J. E.: Theoretical consideration in heat and infra-red detection, with particular reference to the pneumatic detector: A pneumatic infra-red detector, Rev. Scient. Instruments 18: 347 and 357, 1947.
3. Hardy, J. D.: The radiating power of human skin in the infrared, Am. J. Phys. 127: 454, 1939.
4. Hardy, J. D.; Wolff, H. G., and Goodell, H. (ed.): *Pain Sensations and Reactions* (Baltimore: Williams & Wilkins Company, 1952).
5. Henriques, F. C.: Studies of thermal injury: V. The predictability and the significance of thermally induced rate processes leading to irreversible epidermal injury, Arch. Path. 43: 489, 1947.
6. Wormser, E. M.: Properties of thermistor infrared detectors, J. Optic. Soc. America 43: 15–21, January, 1953.

MEASUREMENT OF SWEATING

SID ROBINSON *and* ALINE H. ROBINSON, *Indiana University*

STUDIES OF variations in the over-all rate of sweating, of the distribution and variations in activity of the sweat glands in different localities and of the composition of sweat have involved the development of numerous techniques.

Six general methods have been used for measuring sweating: (*a*) determination of weight loss of the subject, (*b*) collection of water vapor from evaporated sweat, (*c*) direct collection of liquid sweat, (*d*) direct observation by microscope of droplets from sweat pores, (*e*) use of color indicators on skin or by imprints on absorbent paper, and (*f*) measurement of changes in skin resistance to galvanic current. Method (*a*) is used only in measuring total body sweating, but may be adapted for use over both short and long periods of time. Method (*b*) may be adapted for local skin areas or for total sweating. Methods (*c*), (*d*) and (*e*) are useful in the study of activity of the sweat glands in local skin areas. Although skin resistance (method (*f*)) has been used to indicate the activity of the sweat glands, its specificity for the sweating response is subject to question and therefore will not be discussed here.

Quantitative studies of the dissolved substances in sweat have been made by (1) analysis of samples of sweat collected directly from the skin, (2) analysis of sweat residues washed from the skin following periods of measuring the rate of evaporation of sweat from the skin, and (3) estimation of sweat components by difference in material balance studies.

I. Weight Loss Method

Water loss from the total body surface can be measured by observing the weight change of the subject during a timed period and correcting the total weight change for material loss through other channels and material intake. The principle can be applied in continuous observations of water loss and for periodic measurements. Its great advantages are its accuracy and convenience of use, and that periodic measurements of sweating may be made without restricting the activity of the subject. Partitioning of sweating from other exchanges of water by the subject is no more difficult with this method than with others.

1. *CONTINUOUS MEASUREMENT OF WEIGHT LOSS*

This requires a highly sensitive balance such as the Sauter balance* or the more rugged Krogh (25) balance.† In the early part of this century Lombard (32) designed and made his own balance to record continuously the weight loss of human subjects. The Sauter balance is sensitive to about 0.2 g if the rate of weight loss of the subject is not greater than 1 g/min. Its sensitivity is reduced to 0.6 g if rate of weight loss is 5 g/min. In the use of these balances the subject reclines or sits on a cot, chair or ergometer suspended from the balance. The support should be waterproof net or wide-mesh wicker to avoid obstruction of evaporation from large skin areas.

By suspending a bicycle ergometer on the Krogh balance, Nielsen (37) used a kymograph for continuous recording of the weight changes of men during periods of work on the ergometer. In this technique the rhythmic movements of a working man cause corresponding excursions of the writing point, but the general slope of the recording with time shows the rate of weight loss of the subject. Nielsen found the recording to be accurate to about 2 g when subjects were performing light to moderate work on the ergometer. With heavy work it is less accurate.

Adolph (2) described a method of studying rapid changes in the sweating of a man sitting in a light wicker chair suspended on a nonrecording Sauter balance. The subject, whose changes in sweating in warm to hot environments are being followed, should wear only jockey shorts so that the sweat may evaporate from his skin as fast as it is secreted. The subject sits on the balance and is accurately counterbalanced with the balance swinging. The initial readings of the pointer position on the scale and the exact time are recorded. Thereafter the balance is kept swinging throughout the period of observation. The extremes of all swings of the pointer on the scale and their times of occurrence to the second are observed and recorded. The midpoints between simultaneous virtual extremes are found. The rate of evaporative loss of the subject is the difference between 2 midpoints; determinations between 2 swings may err by 0.4–1.2 g, but successive determinations may be averaged to yield more reliable weight decrements. As the man loses weight the balance of the scale is progressively upset and should be periodically brought back into balance by adding weights of 10 or 20 g to the subject's side of the balance. The pointer is thereby displaced by an amount proportional to the added weight, and a new series of midpositions is observed. Midpoints are computed

* Made by August Sauter, Ebingen, Wurttemburg, Germany.
† Made at the Laboratory of Zoophysiology, University of Copenhagen.

at the time intervals of the swing periods (average 30 sec). A succession of body weights is thus obtained at 1 min intervals and, when plotted, slopes may be drawn through them which indicate the rates of weight loss at different times of the experiment.

B. PERIODIC MEASUREMENT OF WEIGHT LOSS

Measurement of change in weight of the body is the only method in use for measuring total body sweat without restricting the subject's activity. The method consists of weighing the subject before and after the period for which his sweat rate is to be determined, adding to his initial weight the weight of all food and water ingested during the observation period, and subtracting from that total the subject's weight at the end of the observation period plus the weight loss by urine and feces and the estimated insensible weight loss for the same period. Accuracy of the method varies with the rate of sweating, length of the period of observation, accuracy of the weighings and accuracy of measurement of the other material exchanges. Sweating of subjects for half-hourly or hourly periods during experiments lasting a few hours may be measured. Daily water exchange over several days or weeks may be followed.

In this method the sensitive Sauter or Krogh balances (mentioned earlier) may be used in studies of insensible weight loss. For periodic measurements of weight loss in men under stresses of work and/or heat, it is more convenient to use an accurate platform balance.‡

PARTITIONING OF SWEAT

A procedure for partitioning sweat from other water exchanges is used when analyses of total sweat residues are made to determine the concentrations of solutes in sweat (44). Before starting the period of observation, the subject is weighed nude (to the nearest 0.002 kg) and dons light clothing. At the end of the period he is weighed nude again. Food and water intake and urine and feces output, if any, during the observation period are measured carefully. If the subject maintains 1 metabolic level steadily throughout the period, pulmonary ventilation and respiratory exchange for 5–10 min during the experiment are determined by collecting expired air (in a Douglas bag) and analyzing it for O_2 and CO_2. If he changes from 1 metabolic level to another during the experiment, the respiratory exchange should be determined in both activities. Wet and dry bulb temperatures of expired air are de-

‡ The special platform balance #1100 made by the Buffalo Scale Company, Buffalo, N. Y., is excellent for this purpose. It is graduated to 2 g and has a capacity of 125 kg.

termined by placing appropriate thermometers or thermocouples in the insulated expiratory valve (9). Wet and dry bulb temperatures of ambient air are determined during the experiment either by a recording potentiometer or by reading wet and dry bulb thermometers at frequent intervals.

The rate of sweating during the observation period is then calculated by

$$S = \frac{(W_1 - W_2) + C - P_v - M - V - I}{T} \qquad (1)$$

where S is the rate of sweating in kg/hr; W_1 = initial weight of subject in kg; W_2 = subject's weight at end of period; C = weight of water and food consumed; P_v = evaporation from lungs; M = weight of CO_2 excreted less O_2 intake; V = weight of urine and feces; I = insensible loss by diffusion from the skin, and T = duration of period in hours. Evaporation from the lungs is the difference between the water vapor contents of inspired and expired air. These are calculated from the volume, temperature and humidity of expired air and the temperature and humidity of inspired (ambient) air. Insensible water loss from the skin should be estimated only in case the skin is kept relatively dry by evaporation of sweat as it is secreted, since there is no diffusion of water outward if the skin is wet. Cutaneous insensible loss may be estimated from the data of Pinson (40), who found it to be 5.6, 7.3 and 9.3 g/m^2/hr at skin temperatures of 30, 33 and 36 C, respectively.

In partitional calorimetry only total evaporation is measured, and it is not necessary to measure cutaneous insensible loss and pulmonary evaporation separately from sweating. Total evaporation by the subject is calculated by

$$E = \frac{(W_{c1} - W_{c2}) + C - M - V}{T} \qquad (2)$$

where E is evaporation in kg/hr; W_{c1} = the initial weight of the clothed subject; W_{c2} = his weight with the same clothing at the end of the period, and the other factors are the same as in equation (1).

In single experiments which are completed within a few hours subjects should defecate before starting in order to avoid the necessity of defecating during the sweat collection period. In laboratory work experiments, water is best administered by having subjects siphon it from an overhead reservoir into 500 cc graduated cylinders, observe the volume and drink it through a tube. In field studies, water is administered in canteens, the volumes of which are accurately measured.

II. Measurement of Evaporation from the Skin

Water loss from the body may be determined by measuring the water vapor as it evaporates from the skin. This is done for the whole body in the respiration chamber (18, 39), in the respiration calorimeter (4, 32) or by the infra-red recording gas analyzer (38). Any of these methods can be simultaneously checked by the weight loss method.

Evaporation from a local skin area may be measured by covering the area with a capsule, passing dry air through it and determining the water vapor in the outlet air. Various forms of ventilated capsules have been adapted for flat skin surfaces (8, 26, 40), hands (17), fingers, toes and pinna (36). Thermocouples have been installed in the capsules for measurement of skin and air temperatures (26, 40). The capsules for flat skin areas are of various sizes to cover from 5 to 20 cm^2 of skin surface and may be sealed to the skin with Duco household cement (12), heavy stopcock grease (40) or rubber cement (8). The capsules for the finger and pinna are sealed into holes in rubber membranes. Air flow through capsules has been produced by a motor-driven blower (26), by suction (40), by passing oxygen from a cylinder (3, 8, 36) or air from a compressor-tank system (17). Moisture in air leaving the skin capsule can be measured by an infra-red gas analyzer (3), or weighed after collection by condensation in refrigerated tubes (8, 36) or weighed after absorption by H_2SO_4 in pumice stone (40), by $CaCl_2$ (26) or by silica gel (17). The hair hygrometer was used by some earlier workers to indicate the moisture content of air from the capsule but is subject to a long lag and considerable error.

The method of collecting and weighing the water evaporated from the skin has the advantages that the apparatus is simple and inexpensive and a number of units can be used for simultaneous measurements of sweating on different skin areas; its disadvantage is that it is more laborious than the infra-red analyzer method and is not adapted for observing rapid changes in sweating since the measurements are periodic.

A. RESPIRATION CHAMBER AND CALORIMETER

The respiration chamber consists of an air-tight cabinet with systems for measuring the subject's O_2 intake and for absorbing and weighing his expired CO_2 and water vapor. The respiration calorimeter measures respiratory exchange and water output and also heat dissipation of the subject by radiation and convection. The chambers which house the subject in this equipment may vary from the minimal size of a coffin to sizes capable of accommodating

a man on a bicycle ergometer. Water given off by the subject is measured by circulating the air from the chamber through sulfuric acid and determining the weight gain of the latter.

The advantage of this system for measuring sweating is that numerous other physiologic measurements on the subject can be made at the same time. The disadvantages are that the equipment is elaborate and expensive, the observations are necessarily periodic and not adapted for measuring rapid changes in sweating, and the method does not directly differentiate pulmonary from cutaneous water loss. Descriptions of this method and equipment (4, 18, 33, 39) will not be repeated here.

B. INFRA-RED GAS ANALYZER§

The methods described here measure the change in water vapor concentration produced in a stream of gas when the gas is passed over the skin. An infra-red gas analyzer was used to measure

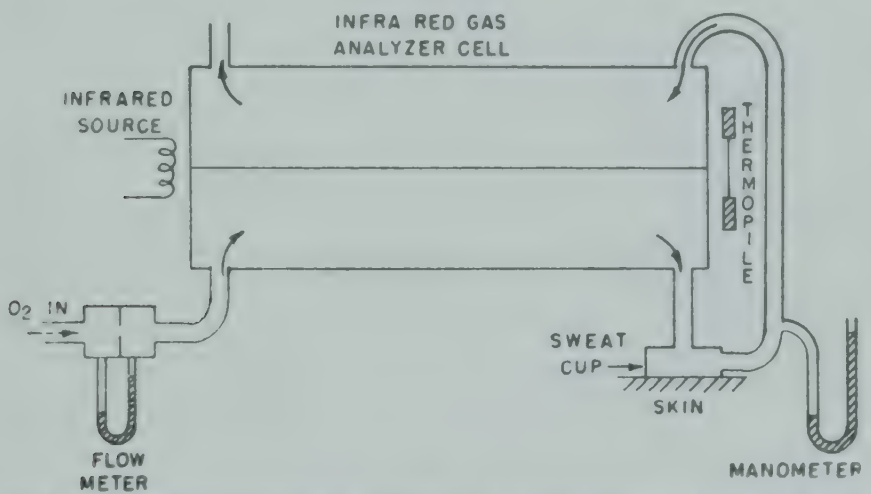


FIG. 1.—Infra-red apparatus for measuring evaporation from small areas of skin.

the difference in water vapor concentration. An N.D.R.C. Selective Gas Analyzer,‖ Model IV (14, 15) was modified (38) for the tests. The measuring cells, original and modified, are described in detail in the reports cited; a schematic diagram of the modified cell is shown as part of Figure 1. The cylindrical cell is divided into 2 portions by a septum, and the ends of the cylinder and septum are sealed to lithium fluoride windows. Infra-red radiation from the heated nichrome source is split into 2 beams, 1 passing through the

§ This portion on the use of the infra-red gas analyzer was written by Dr. E. D. Palmes.

‖ Made by Leeds and Northrup Company; price, $5,000.

reference gas and the other through the test gas. Each beam then hits 1 set of junctions of the thermopile. The potential developed by the imbalance produced is amplified and recorded. With the reference gas passing through both sides of the cell, the thermopile output is balanced by an external potential to give a zero reading. The zero drift is small and linear. The response of the analyzer was

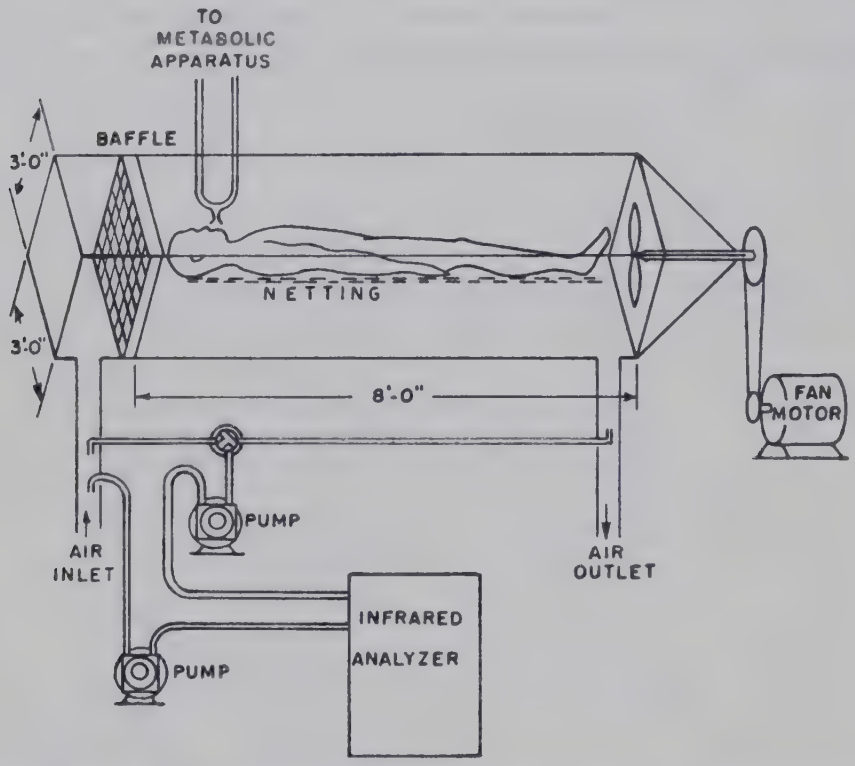


FIG. 2.—Infra-red apparatus for measuring evaporation from the entire skin surface.

practically linear with respect to water vapor concentration in the ranges used.

The analyzer was employed in measurement of evaporation from the whole man (38) and from small skin areas (3). A schematic diagram of the apparatus for measuring water loss from the entire skin surface is shown in Figure 2. The subject reclined on a waterproof netting support and a stream of air from an air conditioning unit was passed over him at a fixed rate. One pump supplied a sample of the inlet air to the gas analyzer and another could be used to sample either inlet or outlet air. The instrument zero was set while inlet air was being compared to itself. The second pump was then connected to the outlet and the difference between inlet and outlet recorded on a strip chart recorder.

The analyzer was calibrated against air of known water vapor

content and the response was found to be practically linear in the range of 0.7–1.7% water vapor. The calibration was not, however, constant from day to day. In a series of 23 calibrations, about ¾ varied 5% or less from the average, and the extremes were about 20%. In all experiments on men, evaporation from skin as determined by correcting weight loss for the entire experiment was considered more accurate than that obtained by using an average calibration value. Thus, the analyzer record was used to distribute the evaporative loss in time. To accomplish this, the entire record for an experiment is integrated by calculating the area under the curve in deflection times time units. This divided by time gives the average deflection for that experiment. Evaporation from the skin is then calculated from weight loss by correcting for urine excreted, water ingested, evaporation from the respiratory tract¶ and CO_2 excess. This value divided by the time gives the average evaporative rate. Average rate divided by average deflection gives the rate per unit deflection, which factor is used to convert any deflection obtained during the run to its corresponding rate.

Chamber volume and air flow were such that about 3 min was required to attain 90% clearance of chamber air. The analyzer response to an instantaneous change in water vapor concentration was 90% in about 8 sec.

It was found in using this method that when the subject was not sweating the evaporative rate was consistently low and constant. In periods of active sweating the rate was highly variable. The clearance time of the man chamber was too long to permit satisfactory resolution of these changes in rate. A procedure was, therefore, sought for closer study of these fluctuations.

In collaboration with Albert (3), an apparatus was set up as shown in Figure 1. Tank oxygen was passed at a constant rate through 1 side of the analyzer cell, into a cup which covered about 20 cm^2 of skin, then through the other side of the analyzer cell. The manometer was used to monitor the cup-to-skin seal. In this apparatus the clearance time of the cup was very short and the limiting portion of the system was the analyzer (90% response in 8 sec). The apparatus was calibrated against water vapor in oxygen. Before application of the cup to the skin, instrument zero was set with the cup held against a dry surface.

The apparatus was used to demonstrate the pulsatile character of sweat rates in human subjects, and it was possible to describe the frequency of the variations in rate. By use of duplicate systems

¶ The measurements of water loss were made as part of thermal balance studies, and the subject breathed into a metabolism unit. Evaporation from the respiratory tract, therefore, was not measured by the analyzer.

it could be shown that in widely separated skin areas the pulses in sweat rate occurred simultaneously; however, the analyzer response was too long to permit adequate observation of the form of individual pulses.

C. CONDENSATION OF WATER VAPOR FROM SKIN

1. *Apparatus.*—In this method (36, 8), dry oxygen from a supply tank is passed through a small capsule covering a skin area. The water from the capsulated skin area evaporates into the

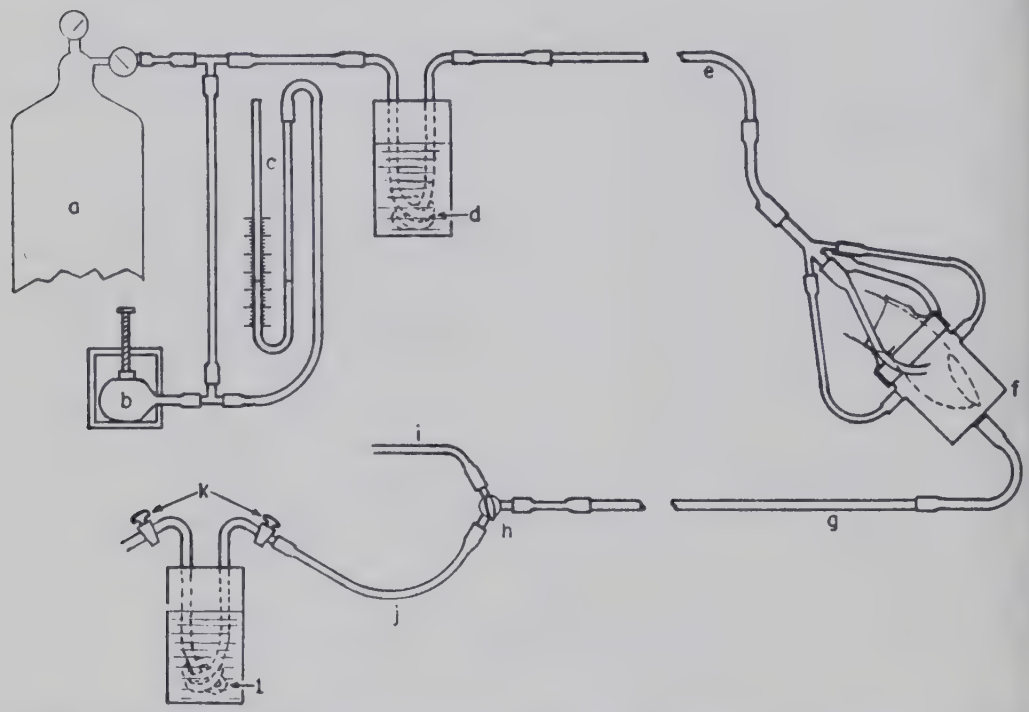


FIG. 3.—Apparatus for condensing water vapor from small skin areas. See text for description of parts. (From Neumann *et al.* (36).)

dry gas and the moisture-laden gas then passes on through a coil of refrigerated aluminum tubing where the water is trapped by freezing. The amount of water evaporated from the skin in a given time is measured by the weight gain of the aluminum coil.

Oxygen is passed from the tank (Fig. 3, *a*), then successively through an aluminum coil (*d*) submerged in a thermos bottle containing a freezing mixture of ethyl alcohol and CO_2 snow for drying, a metal tube (*e*) at least 12 ft long, where it is brought to ambient temperature, the sweat capsule (*f*), metal tubing (*g*), stopcock (*h*), and finally a second refrigerated aluminum coil (*l*), where the evaporated sweat is condensed. A water manometer (*c*) is placed in

he system to indicate gas pressure in the cup during experiments nd to be used in testing the system for leaks. A sweat capsule *f*) made for the finger-tip is illustrated. Ventilated sweat capsules dapted for other skin have been described (8, 26, 36, 40). The luminum collecting coils (*l*) are made of about 1 m of tubing O.D. 4.8 mm, I.D. 3.2 mm) guarded at each end by a metal stop-ock. For convenience in making observations, they are made to weigh 50 g. Simultaneous measurements from other skin areas are nade by arranging for separate gas flows through other capsules nd coils.

2. *Technique.*—To measure sweating, the capsule is sealed to he skin and the system closed and tested for leaks with positive pressure on the manometer. Oxygen flow is adjusted with pressure not exceeding 3.5 cm of water, to evaporate moisture completely from the skin: rates of 100–500 cc/min have been used for insensible loss (36, 40) and up to 800 cc/min for active sweating of skin areas up to 20 cm^2 (26). The gas is allowed to escape for the first 30 min to dry the system. The stopcock is turned and the moisture-laden gas from the cup is passed through the previously weighed collecting coil for the desired period. By use of the stopcock, the flow may be diverted successively through other refrigerated coils to collect a series of samples. The collecting coils are previously prepared by drying inside and out with a flow of room air. They are then filled with dry O_2 at atmospheric pressure and room temperature and weighed on an analytical balance. After collection of water, the stopcocks to the coil are closed, it is brought to room temperature, 1 of the stopcocks is opened momentarily to bring it to atmospheric pressure, and it is weighed again. Weight gain of the coil represents the water loss from the skin area covered by the capsule, and should be related to time and surface area of skin in the capsule.

Measurement of surface area of finger-tip, toetip and posterior portion of the pinna may be made by methods previously described (7, 22). The formula of Isbell (22) for estimating surface area of the finger-tip from its volume is also applicable to the toetip (36).

III. Distribution and Activity of Sweat Glands

In studies of the regional distribution and activity of sweat glands in man, various techniques have been tried: color reactions of sweat with starch-iodine (35), cobalt chloride (45), ferric chloride-tannic acid (46) and silver nitrate (19), microscopic observations of sweat accumulation from the sweat pores (6, 27, 31) and sweat accumulation in a capillary glass cannula placed in a single

sweat duct (26). Recent workers have developed some of these techniques into simple and accurate methods which permit simultaneous observations on the number of active sweat glands in a given area and the quantitative sweat output from an adjacent area. Measurements may be made in a number of skin areas at the same time. From these observations, the average sweat output per active sweat gland may be calculated by dividing sweat secreted/cm²/min by the average number of active sweat glands/cm²/min.

A. COUNTING ACTIVE SWEAT GLANDS

1. *Randall Method* (41).—A dilute solution of iodine, 2–3% in

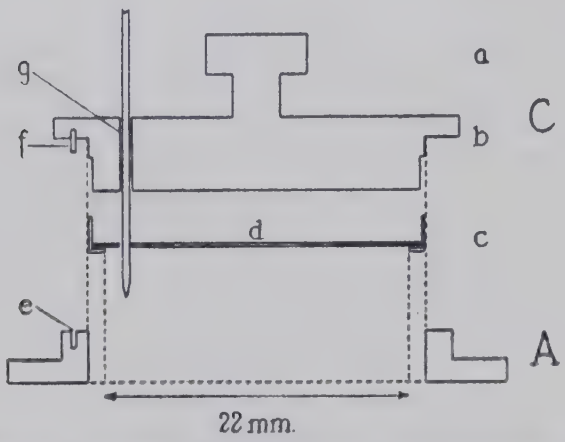


FIG. 4.—Apparatus for sweat gland counts: aluminum skin ring (*A*) sealed to the skin with Duco household cement, and Lucite cover (*C*) used to apply impregnated paper disk (*d*) to the skin inside ring *A*. A stainless steel ring (*c*) holds paper (*d*) onto the cover by friction grip. For orientation of the prints, the pin (*f*) fits into a hole (*e*) in the skin ring, and a needle (*g*) is passed through a small hole in the cover to punch a hole in the paper just opposite the hole in the fixed ring *A*. (From Dole *et al.* (13).)

95% alcohol, is painted on the skin area to be studied, and the alcohol allowed to evaporate. Imprints of active sweat pores in the area are taken on blank pieces of starch-containing paper; no. 13 Voucher Bond was found by Randall to contain enough starch and to have a finish which prevents excessive diffusion in the paper. For an imprint, the skin area is blotted dry and a piece of the paper is pressed lightly over the area for 20 sec; as water coming from the sweat pores dissolves the starch and iodine, blue-black spots appear on the paper. The size of the spots produced in 20 sec indicates the rate of secretion from each sweat pore. These prints may be kept in the dark for several weeks. For permanent records they may be photographed. The number of active sweat glands is determined by counting the spots/cm² on the test paper with a

issecting microscope or from enlarged or projected photographs.

2. *Method of Dole et al.* (13).—Imprints of active glands in a kin area are made on disks of starch-containing bond paper previously impregnated with iodine and cut to exact size by a punch. odine is sublimed into the paper from crystals of iodine in a Petri ish, by spreading 2 or 3 pieces of paper across the top of the dish, overing it and keeping the assembly in an oven at 70 C until the aper acquires a light tan color. The prepared paper must be andled with forceps since it prints water from the finger-tips. For rinting, the prepared disks of paper are applied to the skin during imed intervals in a special sweat capsule (Fig. 4). Iodine reacts

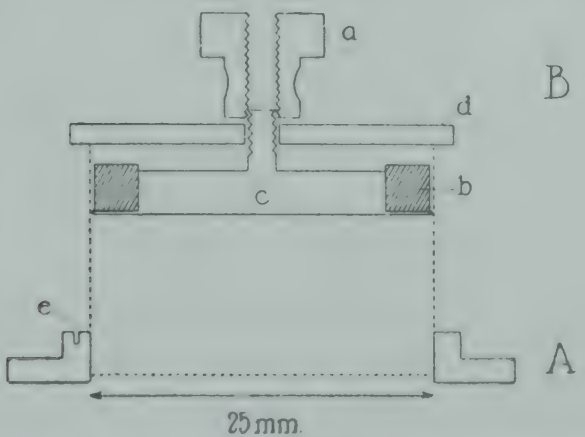


FIG. 5.—Apparatus for collecting sweat: aluminum skin ring (*A*) sealed o the skin with Duco household cement, cover (*B*) and filter paper disks to bsorb sweat. For collection, a preweighed paper disk is laid on the skin nside ring *A*, and cover *B* is sealed into the ring by tightening knob *a* which ulls plate *c* against plate *d*, thereby compressing rubber washer *b* and expanding it against ring *A*. The space between skin and cover is 1 mm deep. From Dole *et al.* (13).)

vith starch in the paper to form a perceptible dot when as little as 0^{-4} mg of water is absorbed from a sweat pore (49). A series of rints of the active glands in an area may be taken during consecuive time intervals by replacing the disks in the capsule.

3. *SWEAT OUTPUT OF A SKIN AREA*

Methods for measuring the sweat output from small skin areas nclude the measurement of water vapor from ventilated skin capules used by Neumann *et al.* (36) and Albert and Palmes (3) lescribed earlier, the method of Dole *et al.* (12, 13, 49) in which he sweat is absorbed by filter paper covered by an unventilated kin capsule, and the desiccated capsule method of Randall (42). n the use of unventilated skin capsules consideration must be given o possible repenetration of unevaporated sweat into the skin dur-

ing collection periods (16). In the ventilated capsule method it is possible to maintain within the capsule an atmosphere approximating that surrounding the rest of the body and avoid accumulation of unevaporated sweat on the skin.

For the unventilated capsule method Dole *et al.* (12, 13) developed a skin capsule (Fig. 5) in which sweat secreted during accurately timed periods is absorbed in filter paper disks placed against the skin under the cover of the capsule. Before a test the required number of filter paper disks (S. and S., no. 589, 25 mm diameter) are placed in glass-stoppered weighing vessels (Pyrex, 30 ml capacity) and each vessel with its contained disk weighed to ± 0.03 mg. All handling of the disks is with clean forceps. The vessels are kept

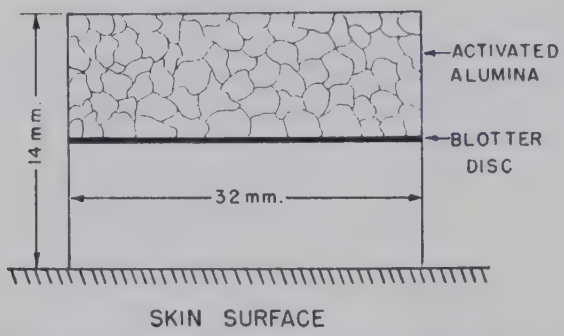


FIG. 6.—Metal capsule inverted over skin during collection of sweat. (From Randall and McClure (42).)

covered with paper hoods and are handled with care to prevent contamination of rim or stopper. The ring and the test area of skin are rinsed with distilled water and dried with analytical grade filter paper. A preweighed disk is set in place and the unit closed. Thereafter, at accurately timed intervals (usually 10–30 min), the disks are replaced throughout the experimental period. If longer collection periods are desired, 2 disks may be used in the chamber at once. On removal of the disks from the sweat chamber, they are placed in their respective vessels, returned to the balance room, allowed to come to temperature equilibrium and weighed again. The weight gain of the disks represents the sweat secretion of the capsulated skin area during the respective periods of collection.

In the technique developed by Randall and Hertzman and their associates (21, 42), the sweat from small skin areas is collected and measured in desiccating capsules (Fig. 6). Each capsule is prepared for use by filling it half full of activated alumina which is held in the capsule by a snug-fitting blotter disk. A second disk of absorbent paper is fitted into the capsule and the unit is covered by a small glass plate. The unit is dried in an oven at 160 C for 3

hr. allowed to cool to room temperature and stored in a desiccator. Immediately before use the whole unit is weighed on an analytical balance. The glass plate and the second paper disk are removed and stored in the desiccator, and the capsule is inverted and taped over the skin area to be tested. At the end of a test period, the capsule is removed and the blotter disk from the desiccator is applied over the skin area to absorb any unevaporated sweat. The disk is returned to the capsule, covered with the glass plate, and the whole unit is weighed again. A blank capsule charged with desiccant is handled similarly except that it is taped to a rubber sheet lying on a board near the subject. The weight gain of the sweat capsule less the weight gain of the blank capsule is a measure of water loss from the skin area. Collections from a skin area during consecutive periods are made by applying a series of desiccating capsules 1 immediately after another.

IV. Collection of Sweat for Analysis

Sweat for analysis may be collected directly by scraping it from the exposed skin into a beaker or test tube (1, 47), by absorbing it in filter paper or absorbent cotton in a capsule covering a skin area (12, 13, 49–51), by allowing it to accumulate in an impermeable glove, sock or bag enveloping a skin region (5, 16, 23, 24, 29, 34, 43, 48) or in capillary micropipets from sweat droplets as they form at the sweat pores (31). If sweat from a capsulated skin area is evaporated as it is secreted and the vapor measured by an appropriate method, the sweat residues may be collected from the skin area and analyzed. For quantitative determination of total body losses of materials dissolved in the sweat, the collection of sweat residues by washing them from the skin and clothing with water is commonly used (20, 30, 34, 44, 52). In long-term observations Dill (10, 11) has estimated materials lost in the sweat by difference in material balance studies without actually collecting samples of the sweat.

A. DIRECT COLLECTION OF SWEAT

Sweat may be collected directly from local skin regions in unventilated capsules or bags covering the skin area or areas. Consideration must be given to errors introduced by repenetration of water into the covered skin and to the fact that lack of ventilation in the glove or capsule makes conditions different from those prevailing normally elsewhere on the body (5, 16, 43). Analyses of sweat collected by scraping it from the exposed skin surface into a beaker or test tube are subject to error due to evaporation. Con-

centrations of solutes in local sweat collected by these techniques cannot be relied on to be representative of the mean concentration for the entire skin surface (1, 5, 16, 28, 34, 43).

1. *Sweat capsule.*—Sweat collected by the capsule method of Dole *et al.* (p. 112) may be analyzed for dissolved substances. After the tests are completed and the weights of the sweat samples determined, the filter paper disks containing the sweat are extracted in the weighing vessels for chemical analysis. Dole *et al.* (12) have described micromethods of extraction and analysis for sodium (flame photometer) and for urea (colorimetric). If sodium analysis is to be carried out, preliminary cleaning of the glassware with acid-dichromate and coating it with silicone are necessary.

From the data, the rate of secretion from the skin area and concentrations of the dissolved substances are calculated. This technique has been applied to studies of local sweat response to subdermal or intradermal injection of a cholinergic drug as well as to sweat response to general stimulation. Simultaneous collections of sweat from different skin regions may be made. Apparatus and techniques for collection of sweat in larger capsules for shorter periods were described by Thomson (50) and Weiner (51).

2. *Impermeable bag method.*—Numerous workers have collected sweat for measurement and analysis from limited skin regions in impermeable bags. This technique lends itself to comparisons of sweat secreted simultaneously by different regions of the body, and large samples can be collected at frequent intervals for macroanalysis. The temperature of different skin regions under study may be independently controlled by artificial means (43).

Bags have been used to collect sweat from various skin regions. Talbert (48) used a rubber bag covering the hand and arm up to the axilla, another for the foot and leg, and a vest with snug fit around neck, upper arms and waist. The lower border of the vest is tucked in all around the waist to serve as a trough for collecting the sweat. Others have used gloves and full-length arm bags (23, 24, 29, 34, 44, 51) and still others, impermeable socks (5, 16).

To remove contaminating materials before the collection period, the skin is thoroughly washed and dried with a cloth previously washed in distilled water. In fitting a bag on arm or leg, care must be taken to exclude sweat from proximal skin regions; for this purpose, a rubber band may be fixed around the upper border of the bag just tight enough to make it snug without constriction. We have taken the added precaution of wrapping an absorbent gauze bandage around the arm and rim of the bag to absorb drops of sweat running down the arm. Accurately timed samples may be

collected without contamination from a glass-to-rubber-tube connection fitted into the lower corner of the bag or in 1 finger of a glove. The tip of the tube is clamped shut except during the collection of samples.

3. *Pipet method.*—Sweat may be collected in a capillary pipet from droplets as they form on the surface at the openings of the sweat pores of palmar skin of the finger (31). Sweat is collected in this way from fields of about 50 pores by direct observation under a dissecting microscope. The fluid droplets are drawn into the pipet by capillary attraction. During periods of collection the finger is enclosed in a glass chamber where water vapor pressure in the atmosphere is maintained at the same level as that on the skin. This prevents loss of water by evaporation from the droplets. Microchemical analyses are made of sweat collected by this method.

B. COLLECTION OF SWEAT RESIDUES

If conditions of air movement, temperature and humidity are such that sweat is evaporated from the skin as it is formed, the nonvolatile solutes left on the skin may be washed off and analyzed. From analyses of residues representing accurately timed periods of sweating and simultaneously determined sweat rates, the concentrations of the solutes in the original sweat may be calculated. With this method, sweat is collected with the skin functioning in its natural atmospheric environment. The technique can be applied most easily and accurately by measuring sweat output by the weight-change method. However, it may also be used in experiments where sweating is measured by the infra-red analyzer.

1. *Total sweat residues.*—The nonvolatile solutes being lost from the body in the sweat should be collected from the entire skin surface under conditions where no sweat is allowed to drip from the skin and care is taken to prevent contamination. This allows the subject to carry on natural activities, such as walking, under a wide range of environmental conditions, with his skin bare or clothed according to the requirements of the experiment. The method used by Robinson *et al.* (44) is described here.

To avoid contamination of the sweat, any clothing, heart leads or thermocouples to be worn by the subject, towels and other objects which he may touch during an experiment should be thoroughly washed and rinsed in distilled water and dried before the experiment. If shoes are required, rubber-soled canvas tennis shoes are probably best for they are easy to clean. During an experiment, other things not listed above which the subject is usually required to touch are noseclip and mouthpiece for respiratory exchange, a

500 cc graduated cylinder and drinking tube for measuring water intake, a rectal thermometer and a waterproof net chair or bed. Any or all of these objects may be washed down with him at the end of the period of collection. Men do not usually have to touch the floor or treadmill except with the soles of their shoes, although as a precaution the hand supports at the sides of the treadmill or ergometer should be washed before the experiment. In indoor treadmill or rest experiments the subject may supply himself with drinking water during an experiment by siphoning it from an overhead reservoir into his graduated cylinder.

Just before the subject is to begin the experiment he washes his entire body with soap in a shower bath and rinses himself thoroughly. He dries himself with a prewashed towel and puts on thermocouples and heart leads if he is to wear them. The period of sweat collection begins at the time he is weighed to ± 2 g standing on a towel on the platform balance. He then puts on the minimal clothing required for the experiment. It is most convenient to wear only shorts, shoes and socks unless more clothing is required for physiologic reasons. We have used a series of hourly periods of collection in laboratory sweat experiments lasting up to 6 hr; Weiner and van Heyningen (52) used 30 min periods in recent experiments. At the end of each collection period the subject removes his clothing and is washed thoroughly and systematically in a bath tub containing 5 liters of distilled water. His clothing, except for his tennis shoes, are washed in the tub at the same time. He dries himself again, is immediately weighed without clothing as before, puts on clothing and begins the next period of collection. The water in the tub is thoroughly mixed, samples are taken for analysis and the tub is washed out in preparation for the next bath. In a work experiment of several hours the subject may walk on the treadmill 50 min out of each hour, 10 min being allowed for the hourly wash-down and weighing. If serial samples are not required, a single collection of sweat residues for analysis may be made for the entire experimental period.

The concentration of each solute in the sweat during each collection period is the amount of the solute in the wash water divided by the volume of sweat secreted during the period. From the subject's weight change, water and/or food intake, urine voided and metabolic and evaporative losses from the lungs, the amount and rate of sweating for each collection period is calculated as described in equation (1).

2. *Sweat residues from limited skin regions.*—Sweat residues for analysis may be collected by rinsing capsulated skin areas following periods of measurement of evaporation by the ventilated capsule

ethod. For this a ventilated capsule of appropriate size consisting f an aluminum ring and a removable lid should be used. When a ransparent lid is used, the capsulated skin area can be observed or completeness of evaporation. Before use, the ring should be ealed to the skin with waterproof cement. Connections for ventilaion should be through the detachable lid. Before the experimental eriod the skin area within the ring is washed and thoroughly dried y blotting with filter paper. The lid is attached and collection of apor begun as already described. Following each timed collection eriod the lid is removed, the skin area quickly rinsed with disilled water to collect the residues, the area dried and the lid ttached again for the next collection period. Collection periods hould be long enough to yield sufficient sweat for accurate nalyses.

Since the skin is kept relatively dry by evaporation during these ollections, sweat secretion is equal to the total vapor collection ess estimated insensible vapor loss by diffusion through the skin. Pinson (40) found that insensible loss from normal skin of men veraged 0.57, 0.73 and 0.93 mg/cm²/hr at skin temperatures of 30, 33 and 36 C, respectively.

C. *BALANCE STUDIES*

The method used by Dill *et al.* (10, 11) allows daily determinaions of both volume and composition of sweat while the subject's ctivity is unrestricted. In Dill's studies the subjects were put on constant diet of known caloric, sodium, potassium, chloride and itrogen content. The diet was maintained during a control period f 10 days in a moderate environment and continued when the nen moved into desert heat. During the control period the subjects' ormal daily sweat volume and urinary output of sodium, potasium, chloride and nitrogen were determined. After the men eached the hot environment, the decrement from control values f the daily urinary output of the significant substances was scertained and assumed to be equal to the increased loss of the ubstances in the sweat. The daily sweat volume was estimated by eriodic measurement of sweat loss (p. 102). The difference beween the daily urinary output of a substance during control and weat periods divided by the difference in sweat volumes for the periods gives the concentration of the substance per unit volume f sweat, provided total daily output equals intake. Several days f work in a hot environment are required for a man to attain salt alance (44).

NOTE.—This section was reviewed by David B. Dill.

REFERENCES

1. Adolph, E. F.: The nature of the activities of the human sweat glan Am. J. Physiol. 66: 445, Nov. 1, 1923.
2. Adolph, E. F.: The initiation of sweating in response to heat, Am. Physiol. 145: 710, March, 1946.
3. Albert, R. E., and Palmes, E. D.: Measurement of evaporative ra patterns, J. Appl. Physiol. 4: 208, September, 1951.
4. Atwater, W. D., and Benedict, F. G.: A respiration calorimeter, Ca negie Inst. of Washington Pub. 42, 1905.
5. Blair, J. R.; Dimitroff, J. M., and Hingeley, J. E.: Studies on foot swe control, Med. Dept. Field Res. Lab. Rep., Sept. 12, 1950.
6. Buley, H. M.: Active sweat glands, a method for their study, Arc Dermat. & Syph. 38: 340, 1938.
7. Burch, G. E.; Cohn, A. E., and Neumann, C.: A method for measuri the area of small irregular surfaces of the human body, Science 9 165, 1941.
8. Burch, G. E., and Sodeman, W. A.: Regional relationships of water lo in man, Am. J. Physiol. 138: 603, March, 1943.
9. Burton, A. C.: Temperature of Skin: Measurement and Use as Index Peripheral Blood Flow, in Potter, V. R. (ed.): *Methods in Medic Research* (Chicago: Year Book Publishers, Inc., 1948), Vol. 1, p. 14
10. Daly, C., and Dill, D. B.: Salt economy in humid heat, Am. J. Physic 118: 285, February, 1937.
11. Dill, D. B.; Jones, B. F.; Edwards, H. T., and Oberg, S. A.: Salt econom in extreme dry heat, J. Biol. Chem. 100: 755, May, 1933.
12. Dole, V. P.; Stall, B. G., and Schwartz, I. L.: Induction and analysis sweat, Proc. Soc. Exper. Biol. & Med. 77: 412, 1951.
13. Dole, V. P.; Thaysen, J. H., and Schwartz, I. L.: Personal communic tion, Jan. 16, 1953.
14. Fastie, W. G., and Pfund, A. H.: Selective infra-red gas analyzer, Optic. Soc. America 37: 762, 1947.
15. Fastie, W. G.; Pfund, A. H., and Peters, C. W.: O.S.R.D. Rep. no. 567 Oct. 31, 1945.
16. Folk, G. E., Jr., and Peary, R. E., Jr.: Penetration of water into th human foot, Rep. no. 181 from Climatic Res. Lab. to O.Q.M.G October, 1951.
17. Forster, R. E., II; Ferris, B. G., Jr., and Day, R.: Heat exchange an blood flow in the hand, Am. J. Physiol. 146: 600, July, 1946.
18. Grafe, E.: Ztschr. f.d. ges. exper. Med. 54: 612, 1927.
19. Gurney, R., and Bunnell, I. L.: Study of reflex mechanisms of sweatin in the human being: Effect of anesthesia and sympathectomy, J. Cli Invest. 21: 269, 1942.
20. Hancock, W.; Whitehouse, A. G. R., and Haldane, J. S.: The loss water and salts through the skin, Proc. Roy. Soc. SB 105: 43, 1929.
21. Hertzman, A. B., *et al.*: A method of partitional calorimetry of individu skin areas, Am. J. Phys. Med. 31: 170, June, 1952.
22. Isbell, H.: The human finger tips: Surface area and volume correlation Human Biol. 11: 536, 1939.
23. Johnson, R. E.; Pitts, G. C., and Consolazio, F. C.: Factors influencin chloride concentration in human sweat, Am. J. Physiol. 141: 57 June, 1944.
24. Johnston, M. W., *et al.*: Hand sweat values in calculation of chlorid and nitrogen balance, Federation Proc. 5: 234, March, 1946.

25. Krogh, A., and Trolle, C.: A balance for the determination of insensible perspiration in man and its use, Skandinav. Arch. f. Physiol. 73: 159, April, 1936.
26. Kuno, Y.: *The Physiology of Human Perspiration* (London: J. & A. Churchill, Ltd., 1934).
27. Kuno, Y.: Variations in secretory activity of human sweat glands, Lancet 1: 299, 1938.
28. Ladell, W. S. S.: Thermal sweating, Brit. M. Bull. 3: 175, 1945.
29. Ladell, W. S. S.: The measurement of chloride losses in the sweat, J. Physiol. 107: 465, 1948.
30. Lee, D. H. K., *et al.*: The effect of exercise in hot atmosphere, M. J. Australia 2: 249, Sept. 6, 1941.
31. Lobitz, W. C., Jr., and Osterberg, A. E.: Chemistry of palmar sweat: preliminary report of apparatus and technic, J. Invest. Dermat. 6: 63, 1945.
32. Lombard, W. P.: A method of recording changes in body weight which occur within short intervals of time, J.A.M.A. 47: 1790, Dec. 1, 1906.
33. Lusk, G.: A respiration calorimeter for the study of disease, Arch. Int. Med. 15: 793, May, 1915.
34. Mickelsen, O., and Keys, A.: The composition of sweat with special reference to the vitamins, J. Biol. Chem. 149: 479, August, 1943.
35. Minor, V.: Ein neues Verfahren zu der klinischen Untersuchung der Schweissabsonderung, Ztschr. ges. Neurol. u. Psychiat. 47: 800, 1927.
36. Neumann, C.; Cohn, A. E., and Burch, G. E.: Quantitative water loss from small areas of skin, Am. J. Physiol. 132: 748, April, 1941.
37. Nielsen, M.: Die regulation der korpertemperatur bei muskelarbeit, Skandinav. Arch. f. Physiol. 79: 193, October, 1938.
38. Palmes, E. D.: Evaporative water loss from human subjects, Rev. Scient. Instruments 19: 711, 1948.
39. Pettenkofer, M., and Voit, C.: Untersuchungen uber den Stofferbrauch des normalen Menschen, Ztschr. f. Biol. 2: 459, 1866.
40. Pinson, E. A.: Evaporation from human skin with sweat glands inactivated, Am. J. Physiol. 137: 492, October, 1942.
41. Randall, W. C.: Quantitation and regional distribution of sweat glands in man, J. Clin. Invest. 25: 761, September, 1946.
42. Randall, W. C., and McClure, W.: Output of individual sweat glands, J. Appl. Physiol. 2: 72, August, 1949.
43. Robinson, S., *et al.*: Effect of skin temperature on salt concentration of sweat, J. Appl. Physiol. 2: 654, June, 1950.
44. Robinson, S.; Kincaid, R. K., and Rhamy, R. K.: Effect of salt deficiency on the salt concentration in sweat, J. Appl. Physiol. 3: 55, August, 1950.
45. Roth, G. M.: Clinical test for sweating, Proc. Staff Meet., Mayo Clin. 10: 383, 1935.
46. Silverman, J. J., and Powell, V. E.: Simple technique for outlining sweat pattern, War Med. 7: 178, 1945.
47. Talbert, G. A.: Effect of work and heat on the hydrogen ion concentration of sweat, Am. J. Physiol. 50: 433, December, 1919.
48. Talbert, G. A.: Hydrogen ion concentration of human sweat, Am. J. Physiol. 61: 493, August, 1922.
49. Thaysen, J. H.; Schwartz, I. L., and Dole, V. P.: Fatigue of the sweat glands, Federation Proc. 11: 161, March, 1952.

50. Thomson, M. L.: Dyshidrosis produced by general and regional ultraviolet radiation in man, J. Physiol. 112: 22, January, 1951.
51. Weiner, J. S.: The regional distribution of sweating, J. Physiol. 104: 32, June, 1945.
52. Weiner, J. S., and van Heyningen, R. E.: Relation of skin temperature to salt concentration of general body sweat, J. Appl. Physiol. 4: 725, March, 1952.

SECTION III

Statistics in Medical Research

ASSOCIATE EDITOR—*Donald Mainland*

INTRODUCTION

DURING THE PLANNING of this Section several workers in different branches of medicine were asked what they thought it should contain; and the differences in the answers were very revealing. One worker said that, after collecting numerical data, he would wish to be able to find here the proper method of statistical analysis. Another said that there was no purpose in repeating what was already available in several textbooks of medical statistics. A third, after listing a considerable number of the commoner techniques, wished emphasis to be placed on the proper design of experiments. A fourth investigator, who has had much experience in giving statistical advice to medical research workers, expressed himself strongly against the "cookbook" type of presentation which, he said, had already done great harm by enabling people to apply statistical tests without thinking, and to apply them with equal ease to data from a good or a bad experiment.

This last expression of opinion must appeal to anyone who conducts research of his own and at the same time gives statistical help in clinical and laboratory investigations in the various fields of medicine. When such experience is supplemented by experience in the assessment of applications for grants in aid of research, in the appraisal of reports on such research and of articles submitted to medical journals, there appears to be a sufficiently broad basis for a generalization—that what most medical research workers need is not, primarily, more knowledge of statistical tests, but a realization of what modern biological statistics implies throughout

the conduct of any type of medical investigation. This need is equally apparent in a feeding experiment on 20 animals and a half-million-dollar survey of peripheral nerve injuries in soldiers.

THE MEANING OF "STATISTICS"

The common conception of statistics as a mass of figures or a set of arithmetical tricks is about as misleading to a medical investigator as the original meaning of the word "artery" (an air tube) would be to a physiologist or surgeon. During the past 30 years, by the close interplay of experimentation and analysis, statistics has come to embody the experimenter's logic, the principles of inductive inference, whereby one tries to draw valid conclusions from experience.

In outline it is easy to see the relationship between statistics as a set of arithmetical techniques and statistics as the principles of experimental design. Having done an experiment, for instance to compare the effects of 2 treatments, each applied to a different sample of animals, we make estimates of the differences in outcome that might occur if chance alone were operating, and we test the observed difference against such estimates. If the difference would rarely occur as the result of chance, we wish to be able confidently to attribute it to the treatment. Obviously, in order to achieve this we must so plan the experiment that, in addition to the treatment, nothing but chance causes differences between the samples. Statistical design shows how to do this. It shows also how to reduce the effects of inherent variation in experimental material, and how to demonstrate the effects of treatments, not only one by one, but in the presence of each other. In brief, it insures not only valid inference but efficient and economical experimentation. Its scope and content are best expressed by a term such as "Design and Analysis of Experiment," the title of a biological statistician's lectureship at Oxford University, or, perhaps better, "Research Design and Analysis," since the word "experiment" has a rather narrow connotation.

Whatever may be its most explanatory title, modern statistics does not claim to be something intrinsically different from the principles and methods of experimenters in general, but to elucidate them and to apply them more thoroughly and more efficiently. Nor does it claim that knowledge of the principles of experimentation, and skill in applying them, constitute research genius or insure epoch-making discoveries in etiology or therapeutics. But very few discoveries are made by haphazard experimentation, and even a genius cannot safely ignore the logic of inductive inference when testing his intuitions.

STATISTICS AN APPLIED SCIENCE

Statistics for experimenters was developed first in agriculture by R. A. Fisher (now Sir Ronald Fisher) and spread thence to animal husbandry and other branches of applied biology. It has been slower in spreading among academic laboratory workers; and this retardation, often by active resistance, has been partly responsible for its delayed entry into clinical research, a field to which it is specially suited, for in origin it is an applied science. That means that those who developed it had continually to face such questions as: "How can I obtain unbiased results under actual working conditions (crop raising, animal breeding, or textile manufacture) in spite of the vagaries of human participants and the variability of animate and inanimate material, instruments and machines?" "How can I, with economy of time, labor and money, obtain results with the precision needed in their practical application?" "What degree of confidence can I attach to my results, and under what conditions?" The applied statistician realizes also that, as in clinical medicine, action has often to be taken without experimental evidence, but he distinguishes clearly between valid experiments and inadequately tested experience, "on which, for lack of anything better, we may have to base our opinions" (1).

STATISTICS AN ART

From the foregoing remarks it can be readily seen that applied statistics is an art as much as a science, and the statistician who is co-operating in an investigation must, as Bradford Hill (2) recently said with reference to clinical trials, be in the project "up to his neck." In laboratory research, after learning the background and purpose, he may be required to show how animals are to be selected, to what cages they are to be allocated, where the cages should be situated, which of the proposed treatments are to be given to certain animals, how effects such as skin lesions are to be measured, when certain animals should be sacrificed, and how observations should be made to insure objectivity.

In a thorough collaboration the distinction between statistician and experimenter tends to disappear, except with regard to their respective techniques, arithmetical, surgical, chemical, and so on. Each learns from the other in hunting for sources of bias and in devising plans to eliminate it.

UNEVEN SPREAD OF STATISTICAL IDEAS

Knowledge of the content and scope of the experimental statistician's task is spread very unevenly in medicine. Many staff members and administrators of medical schools, editors of journals, and

even some heads of foundations still apparently think of medical statistics as concerned largely with vital statistics, hospital records, and other large masses of data, and believe that in experimental work the statistician's function is to apply tests to data already collected. As late as 1951 it was possible for a high official in one of the large foundations for the aid of medical research to assert that a statistician could not show a research worker how to do an experiment.* Less prominent members of the profession can, therefore, hardly be criticized for a similar misconception.

The misconception is symbolized in many schools by the retention of statistics as part of Preventive Medicine or Public Health. Although it is to the credit of workers in that field that for a long time they alone fostered statistics in medicine (and so promoted a critical attitude toward clinical data) this form of association is no longer appropriate, because it entails the subordination of a science which is essentially experimental to a discipline that is, of necessity, probably less experimental than any other branch of medicine.

The increasing tendency for editors to return manuscripts with a request for a standard error or a test of significance may help to focus attention on the need for statistical assistance, but it is a dangerous practice. The proper method, now being adopted by some editors, is to reject articles that do not demonstrate correct design and conduct of the investigation. Even this, of course, could not insure complete reliability, for observation of original worksheets makes one doubtful of many published papers. The papers may seem sound, and may abound in statistical tests, but the real faults, which might invalidate the whole conclusion, may lie buried in the worksheets, or the information that would lead to a detection of the faults may not even have been recorded. This does not imply willful falsification, but simply an unawareness of the risk of bias from certain methods of sampling or observational techniques, or from the rejection of data because they are "out of line."

MEETING THE NEED FOR STATISTICAL HELP

Research workers who appreciate the need for the proper design and analysis of their investigations often ask: "What should I do to secure these features in my work?" "Should I learn the designs and analytical methods myself, or should I always introduce a statistician as collaborator?" "From whom should I seek statistical help?"

With some occasional assistance from a statistician or from an investigator who has had some statistical experience, arithmetical techniques can be learned from textbooks; but even a course of instruction given to a group of research workers will not insure that

* This assertion was a curious echo of one made by an official of similar eminence in the Medical Research Council of Great Britain in 1937, shortly after the first publication of Fisher's *Design of Experiments* (1). Since that time, however, the Council has become a pioneer in the development of the statistically designed clinical trial.

they will be able to design their own experiments correctly in face of the particular difficulties that each experiment presents. Moreover, a research worker should not expect to become familiar with a wide variety of methods, either of design or of analysis. To be a good experimenter, however, he must grasp the principles, and he must be able to apply them to his particular problems. In learning how to do this he will find no safe substitute for personal guidance.

There are, unfortunately, not nearly enough suitably trained statisticians to give more than a small fraction of the guidance that is needed. The supply would increase if medical schools and hospitals would provide adequate financial inducement and rank for statisticians, and if experimentally minded graduates in biology or medicine were not deterred from specializing in medical statistics by the prospect of becoming entangled in vital statistics or hospital records, or of being compelled to take courses in mathematics.

In the foreseeable future it is unlikely that there will be sufficient statisticians to cope adequately with all medical research projects; and a partial solution of the problem might be obtained by recognizing that much research is conducted in order to secure higher appointments and diplomas by graduates whose chief interest is in the treatment of patients. To raise the standard of such research perhaps the best method would be for statisticians, along with clinical and laboratory investigators, to prescribe a limited number of simple techniques of design and analysis, which should be rigidly adhered to unless direct statistical guidance was obtainable.

Those laboratory or clinical workers, however, who are temperamentally good investigators should have access to as much initial guidance as possible. Intimate collaboration with a suitable statistician on 2 or 3 projects often provides such insight into the methods of applying the general principles that a statistician's advice in later projects is largely limited to the answering of specific questions.

Even for this more limited group of investigators the supply of statistical help is yet insufficient. It is, therefore, desirable to use it as efficiently and economically as possible, and the purpose of this series of articles may be described as an effort to promote the economical use of a scarce commodity.

A BACKGROUND FOR STATISTICAL CONSULTATION

The topics that will be discussed are those which the writers† have found repeatedly arising during consultation or collaboration

† Besides sharing in the writing of the articles in which she appears as co-author, Mrs. Lee Herrera contributed criticisms and suggestions to all the other articles in this Section.

with investigators—ideas and general methods which, if appreciated, would provide a background for discussion of specific problems.

Some of the topics may be considered too elementary, others too advanced or too sketchily treated. All that can be said in extenuation is that the selection and treatment are based on the writers' experience with the needs of investigators whose acquaintance with statistical ideas and techniques is very heterogeneous. This heterogeneity exists not only between investigators but within individuals. Some have knowledge of arithmetical techniques, even up to analysis of variance, and yet do not know that randomization is an essential part of an experiment designed for that analysis. Others have a considerable grasp of statistical ideas, even if not expressed in orthodox terms, but find difficulty with elementary points in technique. It may be hoped, therefore, that a fair number of workers may find at least a few items of value in the discussions.

Some of the statements that will be made could doubtless be criticized by statistical purists on technical grounds, but any biological or medical investigator, or anyone who has tried to help others in that class, is skeptical of the value of technical statements that, in order to insure universality of application, are so qualified and complicated as to create bewilderment. He prefers to use an idea or expression which, although not technically exact or complete, enables a worker to draw an inference that is a sound and useful basis for action.

—Donald Mainland.

REFERENCES

1. Fisher, R. A.: *The Design of Experiments* (Edinburgh and London: Oliver & Boyd, Ltd.; New York: Hafner Publishing Company, Inc., 1947).
2. Hill, A. B.: The clinical trial, New England J. Med. 247: 113, 1952.

CHANCE AND RANDOM SAMPLING

DONALD MAINLAND

THE WORD "chance" is used so often in everyday conversation, commonly to imply something unpredictable or with no known cause, that confusion arises when the word is met in its technical sense. This seems to explain in large part why many investigators fail to grasp the basic meaning of statistical estimates and tests, and fail to appreciate one of the most fundamental principles of experimental design, namely randomization.

DEMONSTRATION OF THE EFFECTS OF CHANCE

A definition of "chance" is therefore needed—*the action of a multiplicity of independent causes*—, but this has little meaning unless illustrated by something concrete, and the illustration must be such that the investigator can easily translate it into terms applicable to the material on which he works. For this reason the commonly used illustrations, such as coin tossing and dice throwing, are not as valuable as a disk sampling experiment. In such an experiment a thousand or more metal-rimmed cardboard disks (labeling tags), 1 in. in diameter, are lettered to indicate such therapeutic effects as "improved," "cured," "no change," and "worse," or are marked with figures to indicate measurements. After thorough shuffling in a box, samples of 5, 10 or more disks are taken and their inscriptions are recorded. By replacement of the disks and repetition of the procedure, a large series of samples can be built up.

Such experiments are not only easily visualized by one who has not the time to perform them, but are "real" in the sense that they are in actual use for the experimental study of sampling problems. (Corresponding experiments which anyone can easily perform with random numbers will be mentioned later.) With thorough shuffling the disk experiment comes as close as is humanly possible to the 2 requirements specified in the definition of chance: (1) The factors that determine the presence of any individual disk in a sample are multiple—the various motions imparted by the fingers and by adjacent disks. (2) The factors affecting each disk are independent in the sense that in a thorough shuffling no 2 disks will go through the whole process together, acted on by the same forces to the same extent. Although several may be pushed as a group by 1 motion, a subsequent motion breaks up the group.

RANDOM SAMPLES

Each disk is an *individual*, and all the disks together comprise the *population*. Each sample is a *random* sample, and random samples can be defined as *samples whose differences in composition are determined solely by chance*. The word "random" is often used as equivalent to "haphazard," but there may be much bias in a haphazard sample.

Bias is defined as *something that makes a sample different from what it purports to be*. In a sampling experiment (8) in which colored disks were used it was noted that too high a proportion of disks of a certain color were appearing persistently, and this was found to be due to the paint on 1 set of disks being greasy, so that these disks slipped from the fingers more readily than those of another color. Consequently the samples, which purported to represent the proportions of disks of different colors in the population of disks, were biased. (In terms familiar to experimenters, bias corresponds to systematic error, random variation or random error corresponds to variable error.)

A simple demonstration of human bias is suggested by Yule and Kendall (8). "Ask a friend to recite 'at random' one hundred digits, including zero, and then count the number of odd ones. If the numbers are really random, the number of even ones and odd ones should be about equal, but there will frequently be found a bias one way or the other." The same authors illustrate an analogous bias in scale reading where the terminal digit is supplied by eye interpolation, and the significance of this to laboratory workers is discussed by Mainland (5).

To eliminate doubt, the phrase "strictly random" is often desirable, and when a writer is tempted to use "random" in the colloquial sense he should seek for a substitute. Thus, in a study of the late therapeutic effects of roentgen rays on various skin conditions, the report stated that the patients' records were selected "at random," but the meaning appeared to be "selected without regard to the particular dermatosis." In other cases the implied, but invalid, claim of true randomness has been avoided by substituting such statements as: "The sample was chosen without conscious selection in terms of the severity of the disease." Apart from promoting clarity, such a phrase should make an investigator question the safety of his sampling method.

ARGUMENT FROM SAMPLE TO POPULATION

A disk sampling experiment can be easily translated into terms of clinical or laboratory investigation, for each disk can represent a patient or animal. In all such investigations the argument is from

an observed sample to its population—the series that the investigator would accumulate if he could continue collecting patients or animals of the same kind, under the same conditions, and treated in the same way as was his actual sample; i.e., a population in which the same factors operated to produce variation as operated in the production of the observed sample.

Such a description of the population is not only cumbersome but indefinite. The argument is, in fact, from the observed sample to a population of which it is, or could be, a random sample, i.e., to a population *randomly represented* by the observed sample. Any observed sample could, of course, be a random sample from any one of an indefinitely large number of populations. For example, if the response that is being observed is of the "all-or-none" type (the *A* or *not-A*, or binomial type), a sample of 20 individuals containing 2 *A*'s (10%) might be randomly representative of a population containing 20% *A*'s or a population containing 2% *A*'s, or of any population within this range or outside it (except a population containing zero or 100% *A*'s).

By setting up a disk sampling experiment containing 20% *A*'s in the population (200 disks marked "*A*" and 800 marked "*not-A*"), the investigator could find how often samples of 20 like his actually observed sample of patients or animals (2 *A*'s, 18 *not-A*'s) occurred in, say, 1,000 random samples with 20 disks per sample, i.e., how often his observed sample would occur by chance alone if the true (population) value were 20% *A*'s. He could then do the same thing with disk populations of 25% *A*'s, 30% *A*'s, and so on, until he arrived at a population containing, say 40% *A*'s, from which random samples of 2 *A*'s in 20 would be rare, i.e., they would form only a small proportion of the total samples. Then, adopting a certain "standard of rarity," he could say: "My observed sample would occur so rarely as a random or chance sample if the true percentage were 40, that I believe that it did not come from such a population." In the conventional terminology, he would call his sample of 2 *A*'s in 20 *significantly lower on his specified standard than 40% A's.*

If he accepted 30% as not unlikely, but rejected 40%, then he could narrow the interval by taking other disk populations until he arrived at a population percentage, say 31.7% *A*'s, such that he would exclude all higher values as unlikely, but would accept lower values as not unlikely. The value 31.7%, so chosen, would be his *upper confidence limit;* and by a similar series of experiments, with population percentages below the percentage (10% *A*'s) in his observed sample, he could arrive at a *lower confidence limit.*

When an investigator is not arguing from a single sample to its

possible populations, but comparing 2 or more samples, each treated differently, he asks: "How strong is the evidence that there is a 'real' difference between the effects of the treatments?" This question, also, could be answered by disk sampling experiments. Let us suppose that the data were:

Treatment	Response		Total
	A	*not-A*	
V	2	18	20
W	15	15	30
Total	17	33	50

The hypothesis to be tested is that the 2 samples came from the same population with respect to percentage of *A*'s. The best available estimate for that percentage is obtained by combining the information from the 2 samples, as is done in the lower line of totals, to give 34% *A*'s. In the disk experiment, pairs of random samples (one of 20 disks and one of 30) would be taken and their numbers of *A*'s and *not-A*'s recorded. From a thousand or more such pairs the frequencies of the various combinations would be found, the smaller differences being, of course, commoner than the larger differences; and it would be seen whether differences as great as in the original pair of samples were among the rare differences or not.

Similarly, when responses are recorded by measurement instead of qualitatively, disk sampling experiments could be performed in each case in order to ascertain, in various series produced by random sampling, the rarity of mean differences, differences between means, or other quantities that had been found in the sample or samples of actual measurements.

SUBSTITUTES FOR SAMPLING EXPERIMENTS

A series of sampling experiments after every laboratory or clinical investigation would, of course, be fantastic to contemplate; but because of our knowledge of chance, such experiments are necessary only at the "growing points" of statistical technique. Our knowledge, already adequate for much of the ordinary laboratory and clinical investigation, is based on experience, including experiments of the disk sampling type.

FREQUENCY DATA

In the very simplest cases, common knowledge is sufficient to reveal what would happen if chance alone were operating. For

example, we do not need to perform an experiment to know that, if there were 500 disks marked "*A*" and 500 marked "*not-A*", and we took random samples of 1 disk at a time, replacing each disk after recording it, as we continued sampling we should approach closer and closer to 50% *A*'s and 50% *not-A*'s. By applying the same kind of knowledge, we can see without much difficulty that if we took from the same population random samples of 2 disks we should approach the proportions: Both *A*'s, 25%; *A*, *not-A*, 50%; both *not-A*'s, 25%. Similarly, although with great difficulty, we could work out, for samples of 20 from a population containing 30% *A*'s, the proportions of samples with 0, 1, 2, 3 *A*'s, and so on.

There is, however, no need to work out the results of sampling, step by step for each individual problem, for it has been done in the general form, the *binomial expansion*, which can be used in any particular case. For instance, with samples of 20 and a population of 30% *A*'s, the general form $(p + q)^N$ becomes $(0.3 + 0.7)^{20}$; and by expanding this expression one can find the proportions of random samples to be expected in each class—20 *A*'s, 19 *A*'s, and so on.

Numerous sampling experiments have demonstrated the appropriateness of the binomial expansion, but even this method would be laborious to apply after each investigation. Published tables of the expansion cover only a relatively small range of sample sizes. Therefore an approximation is used, the normal (Gaussian) curve, which simplifies calculation. To compensate for the differences between this curve and the various possible binomial distributions, correction terms have been devised, dependent on sample size and observed percentages (4). The application of these correction terms is, however, rather troublesome to workers who are not using them repeatedly, and therefore tables and graphs of binomial confidence limits have been prepared that demand little or no calculation by the user (5, 6).

COMPARISON OF SAMPLE FREQUENCIES

In the comparison of samples containing frequencies (such as *A*, *not-A*; *A*, *B*, *C*), i.e., in testing the hypothesis that the samples came from the same population, the investigator's need to perform random sampling experiments is removed in a somewhat different way from the one outlined above for single samples. The differences between such samples are conveniently expressed by calculating (by methods shown in elementary textbooks) the quantity *chi-square*, which takes account of the number and sizes of the samples that are being compared. Greater differences between sample percentages produce larger values of chi-square, and therefore larger chi-square values are found more rarely than smaller

values in random sampling from the same population. By using the normal curve as an approximation, it has been shown how often the various values of chi-square would occur by random sampling if the samples were very large, and these values are shown in tables of chi-square probabilities (4).*

For the research worker, whose samples are frequently small, it is important to note the justification for the use of this large-sample technique. Experience with actual samples has shown that the chi-square probability tables can be used for a wide range of sample size, except for certain restrictions specified along with the textbook instructions for the chi-square test. Special attention has been paid to chi-square calculated from fourfold tables (2 samples, each of the *A*, *not-A* type). They have been subjected to extensive sampling experiments and also to tests based on binomial distributions, the "exact" method of Fisher (2). From these tests empirical rules have been derived to show the limits of safety of the chi-square method (5).

For testing the significance of the difference between percentages in 2 binomial samples (a fourfold table), if the samples are of equal size and each of them contains any number of individuals up to 500, the verdict can be obtained directly from published tables (6, 7).

MEASUREMENT DATA

The *t* test for mean differences and for differences between 2 means, and also its extension, the *F* test or analysis of variance for comparison of more than 2 means, are derived from the normal curve; but sampling experiments with actual measurements have shown that these tests can be safely used, even when the frequency distributions of the measurements are far from the normal curve in shape. Provided they are bell-shaped, they can be very much skewed and irregular in outline.

STATISTICAL ANALYSIS AS PART OF THE EXPERIMENT

The foregoing discussion should serve to emphasize an important point. Statistical tests and estimates are not something external or foreign to the experiment that has produced the data. They are simply a completion of the experiment. The tables that show binomial confidence limits, chi-square probabilities, *t* probabilities, and the like, although produced by mathematical methods, are based on experience of the effects of chance and have been shown

* The comparison of 2 percentage frequencies by calculating a standard deviation (standard error) for each percentage can be made exactly equivalent to the chi-square test if certain precautions are taken. For several reasons, including its applicability to the comparison of more than 2 samples (5), familiarity with the chi-square test is recommended.

xperimentally to be appropriate to data of the kind for which hey are prescribed.† They provide the investigator with informaion which otherwise he would have to obtain by an enormous mount of random sampling experimentation.

RANDOMIZATION IN EXPERIMENTAL DESIGN

The ideas of significance testing derived from the random ampling experiment should show why randomization is an essenial part of experimental procedures. This can be illustrated by comparison of treatments V and W on a set of animals. The animals are to be divided into 2 groups or samples; V is to be given to one sample, W to the other. We intend, after the experiment, to apply a test of statistical significance to the results; i.e., we intend to show how often the observed differences would be met if chance alone were operating.

If we find from the test that the verdict is "significant" on our prearranged standard, such as the 5% level, we wish to be in a position to say: "Either the observed difference was due to chance or to the difference between the treatments V and W. But differences as large as the observed difference occur so rarely when chance alone is operating that I am going to attribute the difference to the treatments." Obviously, we must conduct the experiment in such a way that only these 2 alternatives, chance and the treatments, exist; and the only way to do this is by strictly random allocation of treatments to the individual animals.

RANDOM NUMBERS

Random allocation of treatments could be secured by application of the disk sampling method. Thus, 40 disks, each marked with the serial number of 1 animal, could be thoroughly shuffled and then taken out of the box, 1 by 1, without previous inspection of the numbers on the disks. The first 20 disks could be given either the V or the W treatment, as arranged beforehand. A simpler and more generally reliable method, however, is to use numbers that have, as it were, been already thoroughly shuffled and have been tested for randomness. Such are the tables of random numbers of Fisher and Yates (4) and the Kendall and Smith tables, part of which are readily accessible in Arkin and Colton's book (1).

Although Fisher's (3) exposition in 1935 of the need for random-

† Beyond the regions indicated in the text and a few others where sampling experiments have been conducted, there is in medicine a large and expanding territory in which there is needed experimental testing of current methods and of new methods proposed by mathematical statisticians—time-consuming labor that necessitates abundant help at the computing clerk level. Until organizations responsible for the awarding of research grants appreciate the importance of such work, progress will be slow.

ization was quickly appreciated by workers in applied biology, many medical research workers have little or no acquaintance with the method. Therefore introductory instructions on the use of tables of random numbers may not be out of place here.

In all the tables the digits, 1, 2, 3,. . ., 0, are independent of each other, although arranged in blocks for convenience of reading. For application to the experiment on 2 treatments, each applied to 20 animals, the method is simple. Before the experiment starts, allocate the treatments as follows:

Assign to each animal a serial number, which may be simply the order in which a technician will take the animals from the box in which they have been transported. Write down the serial numbers. Decide whether an odd or an even digit from the random numbers shall represent treatment *V* or treatment *W* (counting zero as "even"). Open the table at any page and apply a pointer, such as the corner of an index card, at any part of the page without previously inspecting the numbers. Take the first digit indicated by the pointer and write it alongside Animal no. 1. Move to the next digit—to right or left, above or below, or diagonally, in Fisher and Yates's table; to right or left in the table of Kendall and Smith because it has been most thoroughly checked for randomness horizontally. The first 6 entries might appear thus:

Animal no.	1	2	3	4	5	6
Random digit	7	3	7	6	6	3
Treatment	*V*	*V*	*V*	*W*	*W*	*V*

Having come to the end of a row or column, proceed to the next, and continue either in the same or in the opposite direction.

To avoid allocating too many animals to one treatment it would be convenient, before starting, to make a column of 20 check marks under the letter "*V*" and 20 under "*W*," and then delete a mark whenever a treatment is allotted. When all 20 *V*'s (or *W*'s) are allotted, the remaining animals are to receive the other treatment.

Two people working together can allot treatment to scores of animals in a few minutes. If a number of different experiments are to be performed in the same series, unless all allocations can be done at the beginning it is desirable to note the area of the table used each time, in order to avoid using the same sequence again.

Another simple case is the examination of objects such as histological sections or roentgenograms in random order, to avoid bias due to increasing familiarity, fatigue, a progressive or fluctuating change in an observer's criteria, or changes in the sensitivity of measuring instruments. Write the serial number of each object on a separate index card. Enter on each card a random number in the

rder that they are met in the table. Four-digit numbers, such as 733, 9701, are often desirable to minimize duplication. Arrange he cards in order of increasing magnitude of the random numbers o give the sequence in which the objects are to be examined. If uplicates are found, decide on the order by taking new random igits (e.g., odd and even).

Instructions for more complicated randomization are given by Mainland (5). (For objections to the common method of giving ifferent treatments to alternate patients, see p. 155.)

FACTORS TO BE RANDOMIZED

The general rule for randomization is simple: After systematic ampling (e.g., separation of animals by sex, see p. 140), distribute trictly by chance the effects of all other factors that may possibly ffect the outcome. The application of the rule, however, often equires much thought. Frequently the randomization is either (1) n time, e.g., the random allocation of treatments to patients when they arrive, or to animals as they are removed from a box; r (2) in space, e.g., the random allocation of treatments (or animals) to different cages because of the possibility that difference in ight, heat, and ventilation will affect the animals.

A simple example may illustrate the kind of consideration that should enter into the planning. A box of 30 mice has arrived and we wish to compare 2 treatments (15 mice on each), with 1 mouse to a cage. Even if the animals are similar in weight and other features, we cannot assume that taking them 1 by 1 from the box haphazardly will be equivalent to random selection. For example, possibly the more sluggish (and perhaps, therefore, heavier) mice may be the easiest to reach first. No effort need be made to overcome the bias in the actual removal from the box, and the simplest procedure is to place each animal, as it is taken from the box, into a cage, the cages being filled in any convenient order. The treatments can then be allocated at random to the cages.

In the analysis, the differences between the outcome under the different treatments will be tested against the variation in outcome between animals treated alike. It should be noted, however, that this interanimal variation comprises not only inherent differences between the animals but also any differences that may be due to environment (position or structure of cages). One cannot by analysis disentangle such features as animal weight from the cage differences. To exaggerate, if there were a gradient of weight from the top left cage to the bottom right cage, this might coincide with some environmental gradient such as light or temperature, and an apparent association between response and weight might be entirely

spurious. Moreover, the discovery that there was no significant association between weight and response might also be spurious because the environmental gradient might have counteracted the weight gradient. In reality, the factors generally do not act in simple gradients, but they cannot be ignored, and if analysis of the weight-response relationship, independent of cage effects, is desired it must be built into the experiment.

The simplest way of doing this without subdividing the animals by weights or by groups of cages is to randomize the animals in relation to cages and in relation to treatments separately. If the cages are to remain on the shelves while the animals are being distributed to them, the treatments can first be randomly allocated to (and marked on) the cages and then the animals can be randomly allocated to the cages. If the cages are to be removed from the shelves to receive the animals, they must be replaced in their previously randomized positions.

DESTRUCTION OF RANDOM ORDER

Unless the conception of random order, in space or time or both, is clear, the effect of randomization can be entirely nullified by subsequent steps in the experiment. For example, after treatments (by injection, diet, skin painting, or other method) have been randomly allocated throughout a group of cages, sometimes technicians have, for their own convenience, moved the cages, so that all that were allotted to the same treatment were close together. They should understand that the predetermined cage position is part of the experiment.

Another kind of vitiation of random assignment occurred in an investigation of the effects of medical care. Families that were enrolled in an insurance scheme were randomly divided into 2 groups, *A* and *B*. Group *A* was to receive the ordinary benefits of the plan; Group *B* was to receive, when necessary, additional nursing care, psychological and other guidance. After the random sampling, the families of Group *B* were asked if they would participate in the scheme, and those who declined were eliminated from the group. All families allocated to Group *A*, however, were retained as "controls," to be compared, after several years of the experiment, with Group *B*, in regard to health history.

If randomization or any other procedure in an experimental plan is not carried out properly, or if its effect is nullified by subsequent procedures, the result is more misleading than if the planning had been manifestly unscientific. If, for example, one reads in a report that treatments were assigned to cages strictly at random, one does

ot suspect that the cages were thereafter arranged systematically n groups.

EXPERIMENTS WITH RANDOM NUMBERS

Insight into the effects of chance can be easily gained by experiments with random numbers. For example, imagine that the true (population) frequencies of success and failure after a certain clinical treatment are 70% successes and 30% failures. Call the random digits 1, 2 and 3 "failures" and the remainder, including zero, "successes." Let each digit represent a patient and take, say, 100 samples of 10 patients. It will soon become obvious how far astray a single small sample can lead one, and therefore how important is the habit of thinking of confidence limits rather than of the percentage actually observed.

REFERENCES

1. Arkin, H., and Colton, R. R.: *Tables for Statisticians* (New York: Barnes & Noble, Inc., 1950).
2. Fisher, R. A.: *Statistical Methods for Research Workers* (Edinburgh and London: Oliver & Boyd, Ltd.; New York: Hafner Publishing Company, 1948).
3. Fisher, R. A.: *The Design of Experiments* (Edinburgh and London: Oliver & Boyd, Ltd., 1935).
4. Fisher, R. A., and Yates, F.: *Statistical Tables for Biological, Agricultural and Medical Research* (Edinburgh and London: Oliver & Boyd; New York: Hafner Publishing Company, 1948).
5. Mainland, D.: *Elementary Medical Statistics: The Principles of Quantitative Medicine* (Philadelphia: W. B. Saunders Company, 1952).
6. Mainland, D., Herrera, L., and Sutcliffe, M. I.: *Statistical Tables for Use with Binomial Samples in Medicine and Biology*, in preparation.
7. Mainland, D., and Murray, I. M.: Tables for use in fourfold contingency tests, Science 116: 591, 1952.
8. Yule, G. U., and Kendall, M. G.: *An Introduction to the Theory of Statistics* (London: Charles Griffin and Company, 1940).

THE PLANNING OF INVESTIGATIONS

DONALD MAINLAND

ANY MEDICAL investigator who becomes acquainted with the use of statistics in applied biology or in industrial research cannot fail to be impressed by the wastefulness that he sees in much medical research, in both the clinical and laboratory fields. Modern methods of experimental design are often quickly approved when they are found to save money; for instance, when a factorial experiment, taking only a few hours to outline, saved thousands of dollars in the manufacture of an explosive. But there are more serious forms of waste than waste of money. There is waste of time and there is needless suffering of animals and human beiEgs; and perhaps the greatest waste occurs when an investigation is so conducted that any conclusion that can be drawn from it must inevitably be equivocal.

These remarks are not intended to minimize the rich harvest now being gathered from medical research, but to emphasize the possibility of reducing its cost and of obtaining fewer spurious results. Proper planning does not guarantee a successful experiment, but it makes success more likely; and attempts to draw up a plan are equally valuable if they reveal that a proposed investigation would be futile.

GENERAL PLANNING

Dr. G. W. Beebe of the U.S. National Research Council, while directing statistical work on many large research projects, was so impressed by the waste of time and money due to the inadequate formulation of projects that he drew up an Appendix on Statistical Design, to supplemeEt the questions asked on the standard application forms for grants in aid of research. With his permission it is reproduced here.

Appendix on Statistical Design (Beebe)

Instructions: Experience with the application forms has demonstrated the need for a supplemental statement on the design of the project in most instances. In this appendix you are asked to go beyond the brief statements in paragraph 4 of the application in describing the purpose of your study, the selection of cases for study, the methods of observation, the findings you anticipate, and the methods of arriving at your conclusions. The following questions occur frequently in reviewing applications and

ill serve as useful signposts in formulating your design: in your discussion lease answer those which are pertinent to your study:

1. What questions is the study planned to answer?
2. How will any roster of patients be selected?
3. Do you hope to generalize from the findings on your sample to some much larger group of which your sample is but a part?
4. How did you decide on the size of this group?
5. If controls are to be used, just what is their place in the project?
6. How are they to be selected?
7. If different treatments, etc., are to be compared, how are they to be assigned to control and test groups?
8. Can you exemplify the tabular form of the main findings you anticipate?
9. How do you expect to draw conclusions from these findings?

)ften investigators find it profitable to obtain statistical consultation in leveloping an experimental design. Has such consultation been used in ormulating the design of your investigation?

Although Dr. Beebe's questions were designed primarily for linical surveys, they are broad enough to cover experimental rojects. The rest of this article will be concerned with experiment lanning in more detail. Nonexperimental human surveys are discussed later (p. 159).

PIECEMEAL EXPERIMENTS

An excuse for haphazard investigation is often that it is a preiminary small-scale or pilot study. But even a pilot study should be arefully planned; whereas frequently what may be called "piecemeal" or "partial" experiments are conducted. On a few animals or uman subjects (perhaps even on a single individual) an experiment is performed. In the next experiment, according to what *ppears to be* the result of the first, a dose is changed, and perhaps time interval or some other variable is changed also. In a third xperiment, some other changes of dose are made, operative techiques or methods of measurement are changed, and perhaps nother test substance is introduced; and so on, until, after a dozen r a score of experiments, the experimenter decides to assess his esults. Unfortunately, very little can be extracted from such a eries, even regarding the procedure most suitable for future exeriments or the probable number of subjects required in them. The hole series may, indeed, contain no exact duplicates.

In experiments on a larger scale the same faults may be found, ut correspondingly magnified; or many more subjects may be sed than are necessary; or supplementary experiments have to be erformed in an attempt to eliminate bias that need not have been resent in the main experiment.

RECOMMENDATIONS FOR PLANNING AND GENERAL CONDUCT OF AN EXPERIMENT

The following 12 rules have been formulated during attempts to help investigators, and they may be useful to others.

1. State specifically the *object* of the experiment, which may be to test a *hypothesis* or to form an *estimate*. The hypothesis can usually be expressed as a "no-difference" hypothesis, e.g., that there is no difference in the effect of hormones V and W on blood sugar level. (Note that one can never prove that such a hypothesis is true, but merely try to disprove it. That is the purpose of significance tests.) The object of the experiment, on the other hand, might be to form an estimate of the effect of hormone V on blood sugar level; but in practice, hypothesis testing and estimation are intertwined. For example, to test the above hypothesis there would be required an estimate of the difference between the effects of V and W on blood sugar level.

2. Specify the *population*, i.e., the subjects or material to which your conclusions are to apply, and then keep to that population. (More than one subpopulation may be studied in the same experiment—see (5) below.)

3. Make the plan *simple* unless you are very confident that you can formulate a safe complicated plan, carry it through and secure the necessary analysis of the data. Biological statistics provides a great variety of complicated designs that are very economical of material, time and expense; but in medicine their advantages must be weighed against the risk of breakdown. An intricately balanced scheme can be crippled or even completely ruined by such events as loss of animals or human subjects, delay or omission of observations through pressure of other duties, and an obvious error in a laboratory analysis of material that cannot be obtained again from the same subject under the same conditions. Although methods are available for estimating missing values without introducing bias, they increase the complexity of the analysis, and are generally undesirable for more than 1 or 2 missing items. (This warning against complexity does not imply varying only 1 thing at a time. There are simple plans for testing several factors at once.)

4. Outline the *form of the analysis* to which the data will be subjected. This is the skeleton of the experiment.

5. Perform the desirable *systematic sampling* or *stratification*. For example, the population may be divided into 12 subgroups or subpopulations—male and female animals, mature and immature, of the same species but of 3 different strains. This process not only gives information pertinent to each subgroup but reduces the over-

ll variation that might mask the difference between responses to different experimental treatments. Do not carry this systematic sampling as far as *possible*, but only as far as *convenient* and *useful* n reducing variation and in the application of the results.

Note.—Often a few subjects or types of material outside the main groups are added, such as a few females in an experiment mainly on nales, a few animals of a different species, a few children in an investigation of adults, or a few tests on a drug or preparation different from the ne chiefly used. Usually little or nothing can be concluded from these xtras by themselves, and to incorporate them into the analysis of the nain experiment without bias may be difficult or impossible.

6. Perform the appropriate *randomizations* within each subgroup. They are part of the experiment. Usually an initial one is sufficient, but others should be done whenever there is risk that selection might cause bias. (To omit randomization because one cannot see clearly how bias could occur is like trusting that glassware in chemistry is clean because it does not look dirty.)

7. Be sure that what you think are *replicates* are really replicates, i.e., subjects or batches of material on which all procedures (with the same sequence, time intervals, and so on) have been repeated.

8. Be *uniform* throughout the experiment, e.g., in dose levels and intervals between readings.

9. Unless there is a very good reason, *do not change the plan or omit any item in it* during the course of the experiment. If the plan is manifestly unsatisfactory it is usually better to stop entirely than to continue with fragments.

10. When using the same subject for more than 1 treatment or test, consider the possibility of "carry-over" effects and other *systematic differences;* and design the experiment so as to prevent the resulting bias. (Examples of such effects are mentioned below.)

11. Unless the experiment has been designed and conducted in such a way as to avoid bias, *do not apply a test of significance, or ask a statistician to do so.* Whatever the verdict, whether significant or nonsignificant, it would be dangerously misleading. Such an experiment should be considered a failure.

12. Beware of drawing any *conclusions* from the experiment *except those for which it has been designed.* Any other features, however striking, should be considered as impressions, which may, however, be useful in directing further investigation.

SOME EXPERIMENTAL DESIGNS

No attempt will be made here to enumerate all the useful experimental designs, or even to give a full account of those that are

mentioned. The discussion should, however, show that designs are available to meet various commonly occurring problems. For choice of an appropriate design and application of it, a statistician's help should be sought.

ELIMINATION OF CARRY-OVER EFFECTS

When a glucose tolerance test is repeated more than once in 5 days, the tolerance is likely to be increased, and less marked change is observed with longer intervals between the tests (3). Let it be supposed that the object of an experiment is to contrast the effects of hormones V and W on the glucose tolerance. All patients are first subjected to a tolerance test, and only those whose response is within normal limits are used in the experiment. If there is a carry-over effect from the screening test, the result will not be the same as if that test had not been used, and a carry-over from the test with the first hormone may affect the response to the second. The investigator might, therefore, propose (1) to make a correction for the carry-over, based on records from other subjects, or (2) to allow sufficient lapse of time, again based on other records, to make the carry-over negligible.

Two points should be noted: (*a*) The sample represents only a population of subjects who have received a screening test. (*b*) Corrections based on experiments with other subjects are always questionable; a sound experiment is self-contained. Elimination of the bias must therefore be insured by the design of the experiment, and a *cross-over* (*switchback* or *reversal*) design is appropriate. When the number of subjects to be used in the experiment (an even number) has been decided on, number them serially in the order in which they will receive the first injection. Then allocate at random the 2 sequences, V before W, V after W—an equal number of subjects for each sequence.

Sometimes a carry-over effect is due partly or wholly to a psychological mechanism, as when some unusual or painful process produces less reaction (on pulse rate, blood pressure, etc.) on the second occasion than on the first; or, on the other hand, the second treatment may be preceded or accompanied by an anticipatory reaction from recollection of the first. A psychological reaction was thought to be possible in an investigation of the width of the joint interspace at the knee by body-section roentgenography (2). Investigation of both knees was desired, but only 1 knee at a time could be placed in the central region of the x-ray beam, and the total amount of exposure (of the knee and other regions) to the rays made it unwise to take films of both knees on the same occasion. A second visit was therefore necessary. Differences in muscle tension, however, can change the size of the interspace, and it was thought that on the second visit the greater familiarity with the procedure might cause a difference of tension. (Even if both knees had been x-rayed on the same

casion, one after the other, the same problem would have arisen.) In lf the subjects, therefore, the right knee was filmed on the first visit, e left knee on the second, while in the other half of the subjects the verse order was used, the sequences being randomly allocated in advance.

EQUENCE OF MULTIPLE TREATMENTS

When more than 2 treatments (substances, or doses of the same bstance) have to be applied to the same subject in succession, the mber of possible sequences in which the treatments can be arnged, if each were used, might require too many subjects; for xample, 5 treatments, even if every subject received them in a ifferent sequence, would require $5 \times 4 \times 3 \times 2 \times 1 = 120$ subcts. If there is no risk of carry-over, the design called the *Latin uare* is well adapted for such cases. For 5 treatments, 1 set of 5 atients forms a complete experiment, and as many sets as are esired may be used, with a different pattern in each experiment. he term "Latin" refers to the Latin or Roman letters (*A*, *B*, *C*, c.) representing the treatments: and the term "square," appropriate literally in agriculture where the design was developed, now eans an experiment pattern such as the following, which is one f many arrangements that can be randomly chosen (1).

Patient No.	Sequence of Treatment				
	1	2	3	4	5
(1)	*B*	*A*	*D*	*C*	*E*
(2)	*A*	*B*	*E*	*D*	*C*
(3)	*C*	*D*	*B*	*E*	*A*
(4)	*D*	*E*	*C*	*A*	*B*
(5)	*E*	*C*	*A*	*B*	*D*

t will be noted that each treatment occurs only once in each row nd only once in each column.

In a series of experiments to measure the erythema effect of a betaitting radioactive plaque on 8 areas (about $^1/_2$ in. in diameter) on the osterior aspects of patients' thighs, each area received a different dose exposure time) and the successively increasing doses were given in a roximo-distal order. The investigators disregarded the possibility that roximo-distal differences in vascular supply or skin structure might ffect the response. A Latin square design on 8 patients would have nabled them (1) to explore the possibility of a proximo-distal gradient or regularity in response (comparison of averages of columns, each column epresenting a different position on the skin), and (2) to eliminate such an ffect automatically from the dosage-response relationship.

ULTIPLE FACTORS IN CONTRASTING PAIRS

When more than 1 treatment factor is to be tested and the facors provide contrasting pairs (such as 3 drugs each administered t 2 dose levels), the *factorial* design is eminently suitable. It is uch preferable to the method of varying only 1 factor at a time,

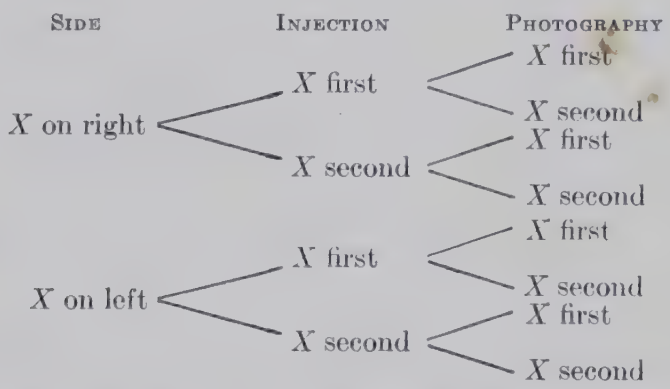
not only because it is very economical but because it enables the interrelationships of factors to be studied at the same time as the main effects.

A factorial design was employed in an experiment to test a substance X, which was believed to have some restraining influence on capillary hemorrhage. In each patient, X was injected subcutaneously in 1 cubital fossa and the same amount of the suspension medium (saline) was injected at the corresponding point of the opposite cubital fossa. Petechial hemorrhages were produced in the forearms by inflating sphygmomanometer cuffs applied around the upper arms, the pressures in the 2 cuffs being as nearly as possible the same.

After a certain number of minutes the 2 cubital fossae were photographed, one after the other, and the numbers of petechiae were later counted "blindfold" in the photograph of each area. The following factors were considered to be possible sources of bias:

1. The order in which the injections of X and saline were made.
2. The side, right or left.
3. The order in which the areas were photographed. Delay in positioning the limb, focusing and loading the camera, and possible interruptions might allow time for a change in the petechiae.

To balance these sources of bias, a scheme of 8 patterns was drawn up:



For each pattern there were 4 patients, to whom the patterns were allocated in strictly random order.

From the data so collected it was possible not only to make an unbiased contrast between the effect of X and saline, but to measure the effects of the factors, singly and in combination. As much information was thereby obtained as would have been acquired by setting up separate experiments to study each factor by itself, using almost as many patients (32) in each experiment.

Computation.—With a little help at the beginning, none of the above-mentioned designs involves very complicated calculation. Indeed, in many types of medical investigation good planning reduces calculation greatly.

INFERENCES FOR WHICH AN EXPERIMENT WAS NOT DESIGNED

Rule 12 (p. 141) contains an important warning against a danger at arises in many experiments when the results are examined—e danger of making comparisons for which the experiment was t designed. In an experiment to test the effect of a vitamin supement on physical fitness, a group of men was divided strictly at ndom into two subgroups, *A* and *B*. Group *B* received the supement and Group *A* did not; but otherwise their diets and living nditions were identical, and all were subjected to the same physil strain and the same tests of fitness. For 6 weeks all had a highlorie diet, and for another 6 weeks all had a low-calorie diet.

The experiment was thus designed to provide a comparison beeen the 2 groups with regard to the *change* in performance when e diet was reduced. The investigators wished in addition, however, to compare the performance on the high and low diets within roup *A* (without supplement) by itself; but here the effect of diet uld not be separated from the effect of time, which could comise factors such as conditioning of the men, boredom with the periment, acclimatization, and change in the climate itself. A gnificant or nonsignificant difference between the performance in e first 6 weeks and the second 6 weeks would, therefore, be open many interpretations.

It was noted that there was much less change in performance hen the diet was reduced than other experiments would lead one expect; but the conditions in 2 experiments are never exactly ike. Therefore the observation should be called merely an *imression*. Such impressions are by no means valueless; and in the ove experiment the impression might provide a rough guide to od requirements in an emergency, if conditions were similar to ose of the experiment. It could also form the basis of another periment designed to estimate the effect of calorie depletion isoted from other factors. The plan for such an experiment would division of the group into 2 samples, 1 to be on the high diet, the her on the low diet; and much more information could be gained after a certain number of weeks, the diets of these 2 groups ere reversed.

REFERENCES

Fisher, R. A., and Yates, F.: *Statistical Tables for Biological, Agricultural and Medical Research* (Edinburgh and London: Oliver & Boyd, Ltd.; New York: Hafner Publishing Company, 1948).

Mainland, D.: Unpublished observations.

Sunderman, F. W., and Boerner, F.: *Normal Values in Clinical Medicine* (Philadelphia: W. B. Saunders Company, 1949).

ANALYSIS IN RELATION TO PLANNING

DONALD MAINLAND *and* LEE HERRERA, *New York University*

IN THE RECOMMENDATIONS for planning an experiment it was stated (p. 140) that an outline of the analysis should be prepared in advance. Here we shall look at a rather common pattern of data and consider questions of planning that arise from it.

In a medical statistics class experiment, capsules supplied by the department of therapeutics were allocated, each in a numbered envelope, strictly at random, 1 capsule to each student. The instructors' list showed which capsules were "*A*" and which were "*B*," and the 2 types were allotted to equal numbers of students; but it was not known until after the experiment that the *A* capsules contained atropine and the *B* capsules lactose. On each student, besides nonmetrical changes such as dryness of the skin and mouth, 6 pulse rates were found, with the following time intervals:

Before Capsule		After Capsule			
15 min	Immediately	15 min	30 min	45 min	60 min

Such data resemble series of readings made in a wide variety of experiments on 2 groups of subjects; for example:

1. Blood sugar levels at hourly intervals on human subjects receiving (*a*) glucose plus a hormone, (*b*) glucose without hormone.
2. Areas of wounds measured at 2-day intervals in animals (*a*) on ordinary diet plus a vitamin supplement, (*b*) on ordinary diet alone, or ordinary diet plus a different vitamin supplement or a different amount of the same supplement.
3. Diphtheria antibody titers at 2-week intervals in premature infants who received antigen (*a*) at birth, (*b*) at 1 month after birth.
4. The albumin:globulin ratios, as an indication of hepatic toxicity of a drug, at weekly intervals in (*a*) patients who have received the drug, (*b*) persons who have received an innocuous substance.

For discussion the outline for pulse rate data can be used as the primary illustration.

OBJECT OF THE EXPERIMENT IN RELATION TO NUMBER OF READINGS

In its most general form the question asked by the experiment was: Do *A*-treated subjects differ from *B*-treated subjects with regard to pulse rate? The hypothesis to be tested is, therefore, that the pulse rate is the same in the two groups—that there is no difference. To answer the question in this form, it would have been sufficient to allocate *A* and *B* at random to different subjects and,

after a certain length of time, find the difference between the mean pulse rates of the 2 groups.

This would have reduced the observations to a minimum (1 reading per subject), but it would be obviously inefficient. In the actual experiment each subject was used as his own "yardstick" and intrasubject comparisons were made. The question actually asked by the experiment was, therefore: Do *A*-treated subjects differ from *B*-treated subjects in their *change* in pulse rate? The hypothesis to be tested is that the change is the same in both groups; and, again restricting readings to a minimum, it would have been sufficient to take 1 reading immediately before treatment and another reading (at the same time in all subjects) after treatment. The mean of the differences between these readings for the *A* sample would then have been compared with the mean of the differences for the *B* sample.

The data actually obtained (6 readings per subject) could provide much more information than a difference between means; for example, an estimate of differences in levels of response of different subjects at different times, and even individual response curves. The question arises therefore: If one does not desire such information should one make so many observations on each subject?

PRETREATMENT READINGS

The number of readings desirable before treatment depends largely on the particular investigation. In an animal blood pressure experiment a continuous recorder indicates when the pressure is stable, and 1 reading immediately before an injection will suffice. If, when readings are not continuous, it is desirable to see whether there is an upward or downward trend before treatment, at least 2 readings should be taken, and the interval must be the same for all subjects. Although there may be no desire to explore this possible trend, it is not uncommon to take 2 or more readings and average them, in order to reduce the pretreatment variation, i.e., to give a more precise "baseline" for comparison with the post-treatment readings. The intervals must, however, be the same in all subjects, for we have no right to assume that the variations are merely random, without trend up or down; and if there were a trend and the intervals were not equal in all subjects, the various "baselines" would not correspond.

POST-TREATMENT READINGS

Even if the investigator does not wish to detect a trend in the readings after treatment, the taking of only one post-treatment reading is often inadvisable because at the time chosen the effect

may not be detectable. It may have already passed off, or it may be yet to come. The experience of a pilot experiment may, however, enable one to reduce the number of readings in later experiments, and also to decide how long the experiment should continue.

As in the pretreatment phase, the times when readings are taken should be the same in all subjects in the experiment, even although the readings are to be merely averaged. It is not essential, however, that the intervals between the readings be the same as before treatment.

UNIFORMITY OF READING TIMES

The main effect of taking readings at different intervals in different subjects can be easily seen. In all analyses, simple or complex, the difference between treatments must be tested by comparing it with the variation among subjects on the same treatment. If, as is usually the case, there is an upward or downward trend in the readings and we take, say, the second reading at 1 part of the slope in 1 subject and at another part of the slope in another subject, we shall incorporate in the intersubject variation some of the variation due to the slope. Consequently, a real difference between treatment effects may be rendered undetectable.

A more serious danger of intersubject differences in reading times, however, is that they may affect unequally the different treatment groups. In such cases the differences due to reading time may be attributed to treatments. For example, if 1 group of subjects is investigated at a time when there is great pressure of other work, there is more likely to be postponement of readings in that group than in a group investigated when more time is available.

Strict randomization of treatments will obviate the risk of this bias; but it cannot remove some forms of bias. For instance, in an investigation of premature infants at different ages (p. 146), there may be a considerable likelihood that examination of the infant, or extraction of blood for testing, will be postponed on younger or lighter infants because of their poor physical condition—a greater likelihood than in the case of older or heavier children. Therefore a bias will be introduced in the contrast between infants treated (e.g. by immunization) at birth and those treated later. Analogous bias can enter into any investigation of sick human beings, and even laboratory experiments are not free from the risk. The only solution, usually, is to discard the data from the subject on whom the observations were made at the wrong time. Then it should be noted that the population represented by the sample comprises only those who are well enough for investigation.

PRECISION OF OBSERVATION TIMES

The precision with which intervals between readings should be measured depends largely on the rapidity of changes in readings, and this cannot be well known before an initial experiment is completed. Minutes are important in an hour's experiment, but in an experiment lasting several months the 9th or 11th day may often be safely taken as equivalent to the 10th day. If there is any possibility of periodic fluctuations in the readings (daily, weekly or monthly), the observation times must be arranged to eliminate bias from that source.

EQUALITY OF INTERVALS BETWEEN READINGS

Careful consideration should be given to the question of equality of interval between successive observations on the same subject. In many long-term experiments it is desirable to explore the early part of the post-treatment period at short intervals, perhaps even daily, but it is impracticable to carry these narrow intervals through several weeks or months. The lengthening of the interval, if done uniformly for all subjects, will not introduce bias into the contrast between treatments *A* and *B;* but it will distort the shape of the response curve.

Even equal intervals, i.e., anything coarser than continuous readings, can fail to outline the response curve properly, and the wider the intervals the greater the risk of distortion. On occasion this can cause misinterpretation of the results. Thus, if the intervals are so wide that they miss the whole neighborhood of the peak regions in the curves, a real difference in the effects of the two treatments may be missed. Or again, treatment *A* may cause a more rapid rise than treatment *B* but both curves may reach the same level. If readings are taken during the ascent of each curve but not near the peak, the verdict may be that treatment *A* causes a higher response than treatment *B*.

SELECTION OF OBSERVATION TIMES

The question of observation times should receive very thorough attention in the planning of experiments. Not only should the foregoing effects of nonuniform and unequal intervals be noted, but the feasibility of choosing certain observation times should be thought of. The possible effects of weekends, holidays, pressure of other work, sickness of subjects or observers, and other sources of bias must be thoroughly considered. It is better to plan on fewer observations if they can be carried out at the proper times than to embark on a scheme calling for many readings and find it unworkable.

METHODS OF ANALYSIS

The main part of the analysis of such data as we are discussing is simple. For each subject, find the difference between the mean of the pretreatment readings and the mean of the post-treatment readings. For treatment A, find the mean of these differences, and likewise for treatment B. Compare these 2 means by the t test. This analysis can, however, fail to demonstrate an effect if at 1 particular observation time there is a difference between the responses to A and B but this is masked by lack of difference at the other observation times. The complete analysis of variance would explore such possibilities.

Workers who are familiar with the t test but not with analysis of variance often explore the differences between readings at different times by applying this test at each observation time and producing such results as:

Post-treatment Observation Time	Significance of Difference between A and B
1	not significant
2	not significant
3	significant
4	not significant
5	significant
6	not significant

Much information contained in the data is lost by this method. Indeed, such tests may even show no significant difference at any individual observation time, whereas analysis of variance, using all the information, may show a pronounced difference between the treatments.

When it is known that, if treatment has any effect at all, it will be in a certain direction (e.g., elevation of pulse rate), the question may be asked: Do treatments A and B differ in the maximal elevation of pulse rate observed under the conditions of the experiment, regardless of the reading time at which the observed peak occurs? In each subject the difference between the immediate pretreatment reading and the highest post-treatment reading could be found and the mean differences for the 2 treatments could be compared. One objection to this procedure can be easily seen by visualizing the complete response curves, as if continuous readings had been taken. With treatment A the true peak might not be near one of the reading times, but might be higher than the peak with treatment B. If, however, a reading time tended to coincide with the peak in B-treated subjects, the average elevation in these subjects might be misleadingly greater than in the A-treated subjects.

The remarks already made regarding the curve shapes when readings are not spaced at equal and very narrow intervals should deter one in most cases from attempting to fit curves, or even from assuming that observed maximal values are indications of real

naxima. The investigator may, however, desire something more han a comparison of mean differences in level of response. Thus, ie may attempt to estimate the time when a peak (or trough) occurs. A common but dangerous way of doing this is to find the ime of the maximal reading for each subject, average these times nd estimate a standard error of the mean. This method treats eading times as if they were random variables, which they are not. Even if readings (Y) are made at narrow equal intervals of time X) and a curve has been fitted, the estimation of X values (the ndependent variate) for particular values of Y (the dependent vari-te) is subject to difficulties because the X values are selected, not andom. In the cases under discussion, with wide and unequal ntervals of X, the estimates are very unsatisfactory. Usually only ery simple methods, such as the following, are justifiable.

Among 20 subjects under treatment A, the observed peak was at the d reading in 2, at the 4th reading in 18. Among 20 subjects under reatment B, the figures were: 11 at the 3d reading and 9 at the 4th. The ifference in incidence can be shown by the chi-square test (or directly rom tables (1)) to be significant at the 1% level. Therefore it can be said hat so far as these reading times revealed the response, there was a ignificant tendency for the B-treated subjects to show a peak earlier than he A-treated subjects.

From binomial confidence limits, used as on page 209, it can further be hown that 2 out of 20 is significantly lower (at the 1% level) than 50%, or the upper confidence limit (99% band) is 38.7%. Therefore there was significantly greater tendency for the peak in the A-treated subjects to ccur at the 4th reading time rather than at the 3d.

REFERENCE

. Mainland, D., and Murray, I. M.: Tables for use in fourfold contingency tests, Science 116: 591, 1952.

THE MODERN METHOD OF CLINICAL TRIAL

DONALD MAINLAND

THE TERM, "modern" in the title of this article implies "developed within the last decade," for the study that can be considered as the pioneer and the model for such studies, the Medical Research Council's investigation (7) of the streptomycin treatment of pulmonary tuberculosis, was reported in 1948. The following discussion of the method is based on that report, the editorial (3) that accompanied it, and articles by Hill (4, 5), Daniels (2), Reid (8), Beecher (1), and Mainland (6), as well as on unpublished personal experience in the planning of such investigations.

The procedure is essentially the rigorous application of the experimental method to tests of therapeutic agents on human patients; and its purpose is to avoid the numerous risks of bias inherent in methods used heretofore, such as (1) the assessment of a new treatment by reference to impressions gained during the use of other treatments, and (2) a comparison of the results actually recorded under a new treatment with those recorded under a previously used treatment.

PRINCIPAL FEATURES OF THE METHOD

The essential features of the method can be summarized under nine headings as follows:

1. *Ethical principles.*—It must first be decided that the trial can be undertaken ethically, i.e., without contravening the moral obligation of the clinician to his patients.

2. *Co-operation.*—The participants must be ready to co-operate wholeheartedly and honestly, carrying out the plan in all its details. That thorough co-operation is possible is shown by the fact that the streptomycin study and later investigations in Britain involved many hospitals, and is further illustrated by the international collaboration (between the United States, the United Kingdom, and Canada) in the study of the treatment of rheumatic fever.

3. *The statistician's role.*—Since the whole scheme is statistical (making allowance for variation and avoiding bias), the statistician must be in from the beginning and throughout the investigation until the final report is ready for publication. As the method be-

comes more widely used, doubtless clinical investigators will assume more and more the role of statisticians who will require to consult professional statisticians only when specially difficult problems arise.

4. *Detailed planning.*—The planning must be minutely detailed, with anticipation of as many possible eventualities as can be thought of. For example:

a) It must at the outset be clearly stated what is meant by a "better" or "more effective" treatment.

b) The population to which the conclusions are to apply must be clearly defined, and in a first trial it is usually best to make the group rather narrow. For example, the Medical Research Council's (7) trial of streptomycin was restricted to "acute progressive bilateral pulmonary tuberculosis of presumably recent origin, bacteriologically proved, unsuitable for collapse therapy, age group 15 to 25 (later extended to 30)." If facilities are adequate, more groups, equally well defined, can be used at the same time. (The uselessness of introducing a few extra subjects, outside the main group or groups, has been mentioned on p. 141.) In a first experiment with any treatment it is desirable to use groups that (i) may be reasonably expected to show benefit from the treatment, and (ii) are not likely to require any special auxiliary treatment.

c) The methods of examination (clinical observations and laboratory techniques) and the persons who will make the examinations must be specified.

d) The observation times must be determined in advance and carefully adhered to. "Every departure from the design of the experiment lowers its efficiency to some extent; too many departures may wholly nullify it. The individual may often think 'it won't matter if I do this (or don't do that) just once'; he forgets that many other individuals may have the same idea" (4).

5. *Prescription of treatment.*—Details of treatment should be thoroughly considered, and made as specific as possible with reference not only to routine or straightforward cases, but to the various complications and emergencies that may arise. This enables the investigators to decide in advance what events during the course of the experiment are to be taken as indications for certain forms of auxiliary or special treatment, and also what events are to be taken as indicating that a patient must, for therapeutic reasons, be dropped from the experiment.

The regimen of the treatment under test (and of the treatment given to the control series), i.e., the scheme of doses and of intervals between courses of treatment, requires special attention. Two procedures are possible:

a) The regimen may be laid down precisely. Thus, except in serious emergencies, the doses and intervals are changed only on certain predetermined indications (a particular temperature or erythrocyte sedimentation rate), and the new regimen to be then adopted is also specified in advance. The objection to this is that it does not represent the method usually applied in the treatment of patients.

b) The change in regimen may be left in the hands of the individual clinician who will decide according to his own judgment of the patient's condition and requirements. The implications of this method should be carefully studied, as in Hill's (5) recent exposition, from which we quote here:

For the fixed-dose regimen the question asked of the trial is of the form: "If to a defined type of patient 2 gr. of drug X are given daily in four divided doses by intramuscular injection and for three months, what happens?" If, however, the physician is free to vary the doses according to his own judgment of the patient's needs as shown by the latter's responses, 2 things must be remembered:

(i) "We have deliberately changed the question asked of the trial; it now runs 'if competent clinicians in charge of defined types of patients use drug X in such varying amounts and for such varying durations of time, and so forth, as they think advisable for each patient, what happens?' "

(ii) "At the conclusion of such a trial we can *in no circumstances* compare the effects of the different regimens of treatment that have been used. These regimens have been determined by the conditions and responses of the individual patients; to observe then, at the end of the trial, the patients' differential conditions and responses in relation to their treatments is merely circular reasoning. . . . .The main danger of this free-for-all trial is the apparently almost overpowering attraction to some clinicians of circular motion."

Even in the simplest form of this flexible regimen, i.e., allocation of a high dose to the more severe cases and a low dose to the less severe cases, it will be seen that we cannot measure the effects of dose per se. The results would have the form:

	Improved	Not Improved
Severe on high dose	...	...
Mild on low dose	...	...

Whatever the differences in proportions (improved and not improved) between the 2 groups, we cannot from such data disentangle the 2 factors, dose and degree of severity of the disease. The only way of contrasting the effects of the 2 doses would be to

take patients with severe disease and randomly allocate the high and low doses in this group, and to do likewise for the mild cases.

6. *Sampling methods.*—The restriction of the experiment to a certain group of subjects (see (4), p. 153) is a *systematic sampling* of the general population, and the process may be carried further in order to reduce the intersubject variation in response, so that the contrast between treatments may manifest itself more clearly. Thus males and females may be separated and the age groups may be subdivided, or in a study of rheumatoid arthritis the patients with spondylitis as well as peripheral joint disease may be separated from those whose peripheral joints alone are affected.

Within each subclass obtained in this way, the treatments (or treatment and control) must be allocated *strictly at random*. There are 2 reasons for this:

a) A logical reason—the provision of a basis for valid inference (pp. 132f.).

b) A psychological reason—the risk of bias, conscious or unconscious, in selection. Even the fear of being biased in 1 direction can produce bias in the opposite direction; and sometimes an effort to make samples alike may make them more alike than would ordinarily occur by chance.

The simplest method of randomization is by random numbers (p. 133). A statistician, secretary or clerk can thus provide the clinician with a set of sealed and numbered envelopes, each containing a slip of paper with the treatment allotted (e.g., "*T*" for treatment, "*C*" for control). When treatment is to start on the 1st patient, envelope no. 1 is opened; no. 2 is opened for the 2d patient, and so on.

The method of "alternation" (allocation of treatment *A* to the 1st patient, *B* to the 2d, *A* to the 3d, and so on) is not advisable. It too readily allows the manipulation of the serial order of patients by clinicians who wish to steer certain patients to 1 or other of the treatments under test. Even thoroughly scrupulous clinicians may be influenced by knowing what treatment the next patient is due to receive. For example, as Hill (4) pointed out: "The method may be insufficiently random if the admission or nonadmission of a case to the trial turns upon a difficult assessment of the patient and if the clinician involved knows whether the patient, if accepted, will pass to the treatment or control group. By such knowledge he may be biased, consciously or unconsciously, in his acceptance or rejection; or through fear of being biased, his judgment may be influenced."

Apart from these risks of psychological bias, the well-known occurrence of trends in the severity of many diseases suggests that

they may occur, unknown to us, in other diseases; and an alternating sequence (*T-C* or *C-T*) of treatment could give a fallacious result if a trend were present.

7. *Objectivity in assessment.*—Every effort should be made to insure objectivity in assessment of the patient's condition; for example, by the devising of simple methods of measurement, such as a graded series of rings to measure the swelling of arthritic patients' finger joints. Objectivity, however, means not only measurement, but an assessment unbiased by knowledge of the treatment that the patient is receiving. In radiological examination this is easily secured, and if it can be obtained in the clinical examination, use can be made of the complex but often very sound "impressionistic" judgment of a patient's condition by an experienced clinician.

Wherever possible it is, of course, desirable to keep the patient in ignorance of the treatment (e.g., by placebo for controls, or by administering 2 treatments that are apparently alike), but it is often impossible to avoid letting the patient know what he is getting, or that he is getting something different from others with the same disease—perhaps a more spectacular technique. Since all diseases have a psychological component, it can be said that 1 element of every treatment is faith, whether engendered by the public press or by the attending physician. Therefore it could well happen that a thoroughly sound comparison of 2 treatments made at a certain time in a certain clinic would give results different from those of an equally sound experiment conducted in a different clinic at the same time or in the same clinic (with the same physicians) a few years later.

This does not, however, nullify or reduce the value of the method. On the contrary, it indicates that it should be applied more widely, e.g., in co-operative experiments at different clinics, and on various types of patients, perhaps differentiated by psychological assessment before treatment. One method of such assessment is the administration of placebos to all patients in a preliminary screening test (1).

8. *Analysis and publication.*—The analysis of the data should be done by an independent observer who is acquainted with the whole procedure; and the published report should be just as detailed and accurate as the report of a laboratory experiment.

9. *Thoroughness of method.*—The method outlined above should be applied thoroughly and completely.

"Such trials are not easy to organize. They demand very careful planning; loose plans, loose methods give loose results, which are just as

equivocal as the impressions of a single clinician and may be more misleading, since a semblance of scientific enquiry is presented. The trials demand close co-operation between clinicians and pathologists of senior standing. They demand trust by these experts in the judgment of an independent person who has had no hand in the clinical treatment of the patients. From such work, no prestige is gained by any single individual. But, given all these things, the results abundantly justify the effort expended. Such methods give, within a year or two years, clear results in a field where unorganized clinical work might not reach a conclusion in less than ten years" (2).

SUPPOSED CONFLICTS BETWEEN THE STATISTICAL AND CLINICAL APPROACH

More and more clinicians are coming to realize that there is no essential conflict between the statistical and the clinical approach to medical problems. But there are still some misunderstandings, and it is desirable to look at them.

1. It is still sometimes said that because every patient is unique it is inappropriate to apply to an individual the results of an investigation that shows "average" effects or "effects in the majority of patients." If this is true, then it seems that "the bottom falls out of the clinical approach as well as the statistical" (5), for it would be illegitimate to build on experience—to apply to a patient a treatment that helped (or seemed to help) a previous patient in a similar condition. Indeed, it is the variability among patients that necessitates a proper method of allowing for it in evaluating treatments, i.e., a statistical method.

Further, after a properly conducted investigation there is nothing to prevent an examination of the individual case histories for peculiar features. Great care must be taken to avoid biased inferences from them, but they may provide useful hints for a further investigation.

2. The assertion that a clinician has no right to experiment on his patients seems to be due to confusion of thinking and perhaps unsavory associations with the phrase "experiments on human beings." The appropriate comment is contained in a medical journal editorial quoted by Hill (5): "In treating patients with unproved remedies we are, whether we like it or not, experimenting on human beings, and a good experiment well reported may be more ethical and entail less shirking of duty than a poor one."

3. A third assertion is that the "scientific" or "objective" approach destroys humanitarianism. In considering this criticism it may be observed that no outlawing of the scientific method would make some clinicians humanitarian; and anyone who works with

good clinicians in the planning of an investigation knows that science does not conflict with humanitarianism. Indeed, it soon becomes obvious that, in order to produce a good plan, giving results that can be applied in therapeutics, the planners must have a sympathetic understanding of patients.

REFERENCES

1. Beecher, H. K.: Experimental pharmacology and measurement of the subjective response, Science 116: 157, 1952.
2. Daniels, M.: Clinical evaluation of chemotherapy in tuberculosis, Brit. M. Bull. 7(4): 320, 1951.
3. Editorial: The controlled therapeutic trial, Brit. M. J. 2: 791, 1948.
4. Hill, A. B.: The clinical trial, Brit. M. Bull. 7(4): 278, 1951.
5. Hill, A. B.: The clinical trial, New England J. Med. 247: 113, 1952.
6. Mainland, D.: Statistics in clinical research: Some general principles, Ann. New York Acad. Sc. 52: 922, 1950.
7. Medical Research Council: Streptomycin treatment of pulmonary tuberculosis, Brit. M. J. 2: 769, 1948.
8. Reid, D. D.: Statistics in clinical research, Ann. New York Acad. Sc. 52: 931, 1950.

CLINICAL SURVEYS

DONALD MAINLAND *and* LEE HERRERA, *New York University*

ANYONE WHO STUDIES the modern clinical therapeutic trial, discussed in the preceding article, is impressed by the contrast in method between the clinical trial and attempts to draw therapeutic conclusions from a survey of clinical records. The difference is noted even in carefully organized surveys, whether conducted in a single hospital or clinic, or by a co-operative study involving a number of centers, or among the personnel of a large industry, or by records and follow-up studies of members of the armed forces. Two questions, therefore, arise:

1. What are the crucial differences between clinical surveys and the modern clinical trial?
2. What role can surveys, in spite of their deficiencies, play in the increase of medical knowledge?

We shall discuss chiefly therapeutic surveys, but shall refer also to surveys conducted for other purposes, such as the study of the etiology or interrelationships of diseases.

In some instances, of course, neither a systematic clinical trial nor an organized survey is necessary to reveal the general benefits of a new treatment, as, for example, when penicillin and cortisone were introduced. Such dramatic differences from previous experience, however, are rare. Moreover, immediately after the discovery of a remarkable treatment, there arise problems of defining more precisely its merits, limitations, and appropriate doses, and of comparing variants (such as different antibiotics) with each other. The 2 questions stated above have, therefore, widespread importance. At the outset we shall consider them chiefly with reference to 2 clinical treatments.

DIFFERENCES BETWEEN SURVEYS AND CLINICAL TRIALS

Defective records.—The most obvious difference between the 2 methods is probably in the quality of records. The ordinary clinical records display incompleteness (e.g., a note on the occurrence of a certain symptom, but no later reference to its persistence or disappearance); lack of information on negative findings; gross errors in laboratory analyses (1), and even fictitious entries (6). All these defects can be found even in records made when a clinician is trying to observe the results of a particular treatment. Arising from pressure of work, from carelessness, or often from ignorance of

the requirements of research data, such defects may not interfere with the proper care of an individual patient, for the clinician may remember the necessary information; but the defects vitiate the records for investigational purposes. Further, one cannot assume that bias would be eliminated by rejecting the imperfect records, for if a clinician's interest is centered on 1 type of treatment or on certain features in a disease, it is very likely that fuller records will be kept of such cases than of others which will, nevertheless, be compared with them in the survey.

Records may be complete for some purposes but not for others; and the clinician who decides to make a survey of cases of a certain disease is not always aware of the type of information that he needs in order to try to answer his questions. As an example, we may consider a physician who wants to study a chronic skin disease in relation to treatments administered at a certain clinic. He traces all available patients that were admitted to the clinic in a 10-year period that ceased, say 3 years ago. He then determines the number of patients who appear cured and the number who still show signs of active disease, and he calculates the proportion of cures.

He fails to realize that the result will be completely meaningless unless he has provided a way of taking into account the length of time elapsed since clinic attendance. This is only 1 of many pitfalls in the search, and anyone who contemplates such a study should first get advice from one who has specialized in the methods of public health statistics, including the drawing of inferences from sample surveys.

In some projects even the most thorough clinical records would not contain the information that is needed. For example, it may be desired to find out something about the frequency with which a disease occurs in the population by age and sex, and it is assumed that the distribution of *patients* by age and sex provides a clue. This is not at all true, for the number of patients in a certain age group, for instance, will depend largely upon the representation of that age group in the population which the clinic serves. To illustrate by exaggeration, if a clinic is set up to serve primarily mothers and infants, one need not be surprised if patients with a certain disease are mainly mothers and infants.

In-clinic records alone cannot provide any information about disease *incidence*. One needs to go outside the clinic to estimate how many potential clinic-goers possess certain characteristics (e.g., how many females aged 20–30). Then one can compute the incidence, namely, the ratio of the number of patients with given characteristics to the estimated number of persons with these characteristics in the actual and potential clinic population. The denominator of the ratio thus comprises clinic patients (with

the specified characteristics) plus the estimated number of persons (with the same characteristics) whom the clinic would have served if they had become ill with the disease under investigation. The mere statement of what is meant by "incidence" shows how difficult it commonly is to form reliable estimates of it.

Comparability of patients.—Serious though the defects in records are, they cannot be considered as constituting an inevitable difference between therapeutic survey data and clinical trial data; for by great effort in a limited survey it would, conceivably, be possible to make the data as complete and accurate as for a clinical trial. The crucial distinction must, therefore, be sought elsewhere—in the comparability of patients subjected to the treatments under contrast.

The gross distortion that can be produced by selection of favorable cases, the transfer of failed cases to other services, as well as other sources of bias in statistics accumulated to demonstrate the merits of a particular treatment, are penetratingly discussed by a surgeon (Ogilvie (9)); but here we are considering the results that might be achieved by efforts to avoid bias and to remove it from data.

(1) It still seems necessary to point out that the surveying of large numbers will not at all guarantee the removal of bias.

(2) Methods commonly used for testing the comparability of patients are often unreliable. For example, the records of each treatment group are classified under headings *A*, *B*, or *A*, *B*, *C*, etc., where the letters can represent any one scheme of classification, such as males and females, children and adults, duration of disease before treatment, presence or absence of a concomitant lesion, presence of a certain symptom in various degrees. To take a simple case, one can imagine 2 groups, *V*-treated and *W*-treated patients, each group containing 100 subjects. They might be exactly alike in sex ratio (50 males and 50 females in each group) and exactly alike in ratio of adults to children (50 : 50 in each group), but the composition might be as follows:

	V-TREATED			*W*-TREATED		
	Children	Adults	Total	Children	Adults	Total
Males	25	25	50	10	40	50
Females	25	25	50	40	10	50
Total	50	50	100	50	50	100

Insofar as men differed from women in their response to treatment (while boys and girls did not differ, or not to the same extent as adults), a comparison of the results of treatment in the 100 *V*'s taken as 1 group and the 100 *W*'s also taken as 1 group could be very misleading. The result could be a verdict in favor of *V* (or

W) even if there was no difference between them; or a true difference might be masked.

In actual analysis, of course, the material is commonly divided by sex, age, and a number of other features, such as physician or clinic, and duration of disease before treatment, that are considered to be relevant to the inquiry, and the comparison of treatments is then made separately on each of these subsamples. This leaves unexplored, however, the numerous other disproportionalities, similar in type to that shown for sex and age; and any 1 of these may introduce bias.

Reduction of bias by repeated subsampling.—The only possible way of reducing such bias is to divide up the data according to all the criteria of classification and obtain a large number of pairs of samples. The members of each pair are then alike in all features of which there is record, except that 1 of them was treated by *V*, the other by *W*. This necessarily entails discarding data from some patients because corresponding *V*-treated (or *W*-treated) patients are not available to make a pair of comparable samples.

The frequent criticism of this procedure, that the samples are too small to show any significant difference, is not strictly valid, because to each small sample a significance test can be applied, and the results of the tests can be combined. The fundamental question, however, is: Would this procedure entirely remove the risk of bias? The answer is: No; because the treatments were allocated by conscious selection, and this is seldom or never equivalent to randomization.

This statement may be elucidated by considering a few of the situations in which treatment is allocated. For example, if during the period under survey a clinician was using *V* on some patients, *W* on others, there must have been something in the particular patient or in the attendant circumstances that made him choose 1 treatment instead of the other for each patient, even though he has no record or recollection of the factor responsible.

Again, if *W* was always given to patients who had failed to respond to *V*, the condition of a patient when he received *W* was not strictly the same as when he received *V*, even though the record shows no difference.

If *V* and *W* were in use at different parts of the period under survey, there is a very considerable risk of difference between patients who, according to the records, were apparently comparable—differences in nursing care, auxiliary treatment, or other features. As an example we may consider a survey of certain types of skin tumors over a 25-year period, where an older method of treatment is to be compared with a more recent one. Besides

possible differences in efficacy of treatments, a host of other variables will enter into the comparison. There may even have been changes in criteria for diagnosis, such that some of the earlier cases, actually misdiagnosed by present standards, are included in (or excluded from) the series treated by the older method. Secular changes may have taken place in the severity of the disease as well as in the population under survey. Distortions of the differences between treatments in such a situation can be caused also by a high proportion of losses in the early cases (incomplete records) and by uncertainty about permanence of cure in the more recent cases.

It should be even more obvious that no valid inference regarding regimen (doses of a particular drug and intervals between doses) can be drawn from a survey of clinical records, because such inferences cannot even be drawn from proper clinical trials if the clinician has prescribed the regimen according to the responses and apparent needs of the patient (p. 154).

It will further be seen that attempts to relate outcome to initial conditions (age, complications, etc.) are also subject to bias when treatment is given at the discretion of the clinician.

Lost cases.—Reference has already been made to the difficulties that the time element introduces into clinical surveys. Since treatments are not allocated at random, differences in the effects of treatments are "confounded with" (confused or inextricably mixed with) many differences associated with time. There is, however, 1 problem connected with time that is present both in the clinical survey and in the clinical experiment, and that is the problem of follow-up. In the course of follow-up, especially if extending over years, it invariably happens that patients are lost. There are losses by death—deaths due to the disease under investigation as well as deaths from other causes. There are losses due to the mobility of patients. Some persons refuse to present themselves for a check-up because they feel completely well and believe that a check-up is unnecessary. Others may have been so dissatisfied with the treatment that they have sought help elsewhere or given up trying to obtain help.

Obviously, every attempt should be made to persuade people to come in for the check-up. More than that is needed, however—an effort to determine as accurately as possible why persons failed to attend. If they have turned for medical assistance elsewhere, an attempt should be made to obtain the necessary information through those channels. Finally, one must realize that assumptions, either explicit or implicit, regarding lost cases are always made in analyzing follow-up data; and if these assumptions have no basis

in fact, the entire results of the investigation may thereby be invalidated.

It will be evident that the task of obtaining information on lost cases is difficult and expensive. It is, however, a less serious problem in a proper clinical experiment than in a clinical survey. Even after an experiment we cannot, of course, assume that the percentage of cures will be the same in those who are lost as in those who are available for study. For example, there may have been a higher proportion of losses among the older patients than among the younger, and the older ones may have been predominantly unsuccessful cases. However, if the treatments were allocated at random, the risk of this bias will have been equalized for the 2 treatments, and therefore it is justifiable to compare treatment effects on the patients that remain available. This is true of all risks of bias due to loss produced by factors that are in no way associated with differences between the treatments.

On the other hand, randomization does not take care of situations in which the treatments, regardless of whether they produced the same proportion of cures or not, differed in such a way as to influence the proportion of lost cases. For example, treatment V and W might be equally efficacious, but V might be more disagreeable or more expensive than W, and this might cause a relatively high proportion of the V-treated patients, now without evidence of the disease, to avoid responding to the request for a check-up.

Contrast with randomization.—Readers who are not very familiar with the implications of randomization may ask: "Why, in the clinical trial, is attention not paid to the numerous possible factors referred to in the foregoing discussion as having a possible influence on the outcome of treatment? Why can one, after strict randomization, say so definitely (p. 133), 'Either chance or the treatment'? Could the apparent superiority of treatment V not have been due to 1 or more of the factors that have been disregarded, or to some factor, at present unsuspected, that may be discovered in the future?"

The answer is implicit in the phrase that is quoted. By it we mean that a strictly random allocation may, indeed, have allotted treatment V to a high proportion of the patients who possess 1 or more of the factors promoting recovery, and may thereby have led us to believe that the treatment itself was responsible. But we guard against too frequent error of this kind by our test of significance. Thus, using the 5% level, we can say that, since randomization produces such an allocation in less than 5% of experiments, we consider it unlikely to have occurred in our particular experiment.

INFERENCES REGARDING DISEASE INCIDENCE—BERKSON'S FALLACY

It is widely recognized that surveys of hospital records and other such data give a biased picture of the relative incidence of diseases in the general population of sick persons, but an important corollary of this, demonstrated by Berkson (2) in 1946, has not been adequately appreciated.

Berkson's demonstration had reference to the search for a possible association between diabetes and cholecystitis. There was such a strong impression of the existence of this association that some surgeons were removing the gallbladder in the treatment of diabetes. To test the soundness of this belief, the incidence of cholecystitis in diabetic patients was compared with its incidence in persons who came to the clinic for eye testing, because it could not reasonably be suspected that there was any association between cholecystitis and errors of refraction. The frequency of cholecystitis was found to be higher in the diabetics by an amount that was statistically significant; but Berkson showed that such results could be entirely fallacious under 2 conditions that must very often exist:

1. That the occurrence of 2 disorders in the same person gives him an increased probability of admission to a hospital or clinic.
2. That the persons with the disorders under investigation are not represented in the hospital or clinic population in the same proportions as in the general population.

The same kind of bias may affect other comparisons made from clinical or autopsy records; it can, for instance, create an apparent difference in the incidence of heart disease in 2 occupational classes. The bias can also prevent the detection of a real association.

This risk of fallacy, important not only to clinicians but to pathologists and to investigators of human physiology, does not depend on complex biological phenomena, and indeed it is not confined to medicine but applies to any kind of investigation that involves sampling. As Berkson (2) pointed out: "the same results.... would occur if the sampling were applied to. . . .cards instead of patients."

To grasp the principle of the fallacy, it is best to start with a very much simplified example. The following figures, used in a discussion of autopsy data (7), will serve for hospital data also.

Let us suppose that there are 2 diseases, A and B, and that we wish to find from hospital records whether another disease, X, is more common in persons with A or in persons with B. Let us further suppose that the percentage frequency of X in the general popula-

tion is exactly the same, 10% in A patients and in B patients. For further simplicity we suppose that the only diseases in the population are A, B and X, and that there are equal numbers of A's and B's, 1,000 of each in the general population. There are, therefore, 100 A's who have X and 900 A's who do not have X; and the same numbers apply to the B's, i.e., 100B,X and 900 B, *not*-X.

We now suppose that, as is almost universally true under actual conditions, in the hospitals concerned the diseases differ from each other with respect to *admission rate*, which is defined for disease A as: No. of A's admitted $\times$ 100/Total No. of A's in the population. We take the admission rate for disease A as 50%; for B, 20%, and for X, 40%. We now calculate how many patients of each group will be found in the hospital records.

Group A,X.—Total patients = 100. Fifty per cent of them, i.e., 50 patients, are admitted because they have disease A, leaving 50 outside. Of these latter, 40%, i.e., 20 patients, will be admitted because they have disease X. Total admissions of A,X = 70.

We may note that, of the 50 who were admitted because of A, 20 would otherwise have been admitted because they suffered from X; or we can even think of A and X as bringing the patients to the hospital simultaneously, as "multiple causes" of admission. In any case, the total number of admissions (70) is not affected; nor is it affected if we count first the 40 who are admitted because of X and then take 50% of the remaining 60 patients, to give 30 admissions because of disease A.

Group A, *not*-X.—Total patients = 900, of whom 450 are admitted because they have disease A.

Group B,X.—Total patients = 100. Twenty per cent, i.e., 20 patients, are admitted because of B, leaving 80 outside. Of these latter, 40%, i.e., 32 patients, are admitted because they have disease X. Total admissions of B,X = 52.

Group B, *not*-X.—Total patients = 900, of whom 180 are admitted because they have disease B.

In summary, the following patients will be found in the records:

	X	*not*-X	Total
A	70	450	520
B	52	180	232
Total	122	630	752

The percentage frequencies of X are as follows:

Of the A's, 70 $\times$ 100/520 = 13.46% have disease X.
Of the B's, 52 $\times$ 100/232 = 22.41% have disease X.
Difference = 8.95, or approximately 9%.

In an actual investigation we should ordinarily test the significance of this difference; and when this is done here, chi-square (with Yates's correction) is found to be 8.81, above the value (6.635) required for significance at the 1% level. And yet we know that in the parent population the percentage frequency of X in the A's and the B's was exactly the same. The fault is not in the statistical test, for such tests never claim to do more than show how often certain things would occur if chance alone were responsible. Something more than chance was operating here, but it was not a closer association between X and B than between X and A. It was a bias in the sampling, due to the lower admission rate of B, which, among patients in the hospital, causes one to find a higher proportion of the B patients afflicted with X than is found among the A patients. The phenomenon can be described as a competition among admission rates. The rate in A offers stronger opposition to X than does the rate in B; i.e., more of the B's are admitted because of X and fewer B, *not*-X than A, *not*-X.

The best way to become familiar with Berkson's fallacy is to substitute different numerical values in the foregoing example. Trying to visualize, next, the great complexity of competing rates that must occur under actual conditions, one becomes aware of the great doubt that must attach to inferences from hospital records regarding the relative incidence of different diseases, and hence regarding any conclusions regarding the etiology or interrelationships of diseases that may be drawn from such records of incidence. (A discussion of the complexity of factors affecting autopsy data (7) is largely applicable to clinical records also.)

Just as the clinical trial is being substituted for surveys in therapeutic investigation, so are more reliable methods coming to be used instead of autopsy incidence and hospital incidence data for the study of the etiology and interrelationships of diseases, both by pathologists and clinical investigators. Among the chief methods now being developed is intensive longitudinal study of population groups. The techniques and difficulties of such methods are well illustrated by papers from the Milbank Memorial Fund Conference (8) held in 1951 and in articles by Cochran (3), Cornfield (4), Hansen and Hurwitz (5).

USES OF SURVEY DATA

The great risks of error in the survey method of therapeutic and etiological research lead to the second question at the beginning of this article: What role can surveys play in the increase of medical knowledge? Their uses can probably be summarized under 4 headings:

1. To provide information for administrative purposes—a wide variety of information ranging from the number of operating chairs required in a dental clinic to an estimate of premiums required to finance a scheme of medical care.
2. To obtain etiologic information when experiments are impossible.
3. To assemble and clarify therapeutic experience when a proper clinical trial is not feasible.
4. To obtain data that may be useful in designing a proper clinical trial or other experiment.

Regarding the last 2 uses of surveys, the following 3 points should be emphasized:

a) When a new treatment appears, it is very unwise to postpone a proper trial. An imperfect experiment or a survey may give such a favorable impression that a proper trial will be difficult or impossible to organize. On the other hand, an imperfect test may give spuriously negative results and cause a useful treatment to be dropped.

b) A proper trial should not be too readily dismissed as unfeasible. Information from 50 patients in such a trial may be more valuable than information from 500 patients in a survey of clinical experience.

c) Since the risk of bias can never be eliminated, a survey can do nothing more than provide information on which, for want of anything better, the clinician has to act, either therapeutically or in planning a proper investigation. Therefore it is wasteful to push the analysis of survey data beyond the point where it will provide some directly useful information.

For example, from a survey of rheumatic disease clinics a physician might wish to know rheumatologists' impressions regarding the value of cortisone in advanced ("burnt out") cases of rheumatoid arthritis. If he wished to act on this information when treating patients, he would want to know the rheumatologists' definition of "advanced" or "burnt out," but he would be unlikely to wish to divide such cases into narrow age groups, or according to the particular peripheral joints involved, or according to the frequency of subcutaneous nodules. If he were contemplating a proper clinical trial and wished to estimate the number of cases that might be required to demonstrate an effect of a certain magnitude, he likewise would gain little or nothing by fine subclassification of the data obtained from the survey.

The very natural desire to find out as much as possible from the data should be tempered by the knowledge that every additional question will add to the time and cost of the survey. One more

question may add several days to the time required for card sorting and analysis. Much time and money spent on surveys could be more profitably devoted to proper experiments.

PLANNING A SURVEY

When a survey of clinical experience is found necessary it should be continually borne in mind that, so far as it pertains to therapeutic effects and etiological factors, a survey is little more than an opinion poll and a study (uncontrolled) of the behavior of doctors and other personnel. The planners should keep before them a statement of objectives and limitations, for this may do something to prevent waste of effort and, perhaps, to reduce the tendency to invalid inferences after the data are collected. As an example, by no means perfect, of an enumeration of aims, there is shown here the draft of a statement prepared for a Rheumatism Association committee that was planning a survey of rheumatologists' experience with cortisone in rheumatoid arthritis.

Cortisone Survey—Suggested Outline of Objectives

1. General objective—to obtain information that will enable the Association (or other group) to decide whether a controlled comparison of various treatments of rheumatoid arthritis is feasible, and if so how it can best be planned.

2. To ascertain the experience of certain clinics (selected by such and such criteria) during a certain time period, in the treatment of rheumatoid arthritis by cortisone; the data, recording the physicians' findings and opinions, to be expressed after appropriate statistical treatment in terms such as "percentage improved," "percentage not improved," "percentage with side effects," etc. *Not* to assume that these findings were true for other clinics or physicians at the same time period, or applicable now or in the future. (One reason for this: the change in clinical judgment during the last 3 years regarding suitability of cases for cortisone treatment.)

3. To look for differences in response that may *appear* to be associated with certain features, such as age, sex, pregnancy, duration of disease before treatment, different concomitant treatments. *Not* to assume that such appearances are anything more than a hint of a possible association. (The hint may be misleading because of the numerous factors that such a survey cannot detect, and cannot compensate for by a random sampling process before treatment.)

4. To look for differences between the statements of different clinics (and different physicians) regarding outcome; and to seek for hints of possible explanations. *Not* to pursue this search very far, because of the numerous undiscoverable factors undoubtedly present, which could introduce bias in either direction (apparent association or apparent lack of association). (The chief use of Item (4) would perhaps be to convince the Association that a proper trial was necessary.)

5. To form an impression of the facilities of the clinics, and their suitability, for possible co-operation in a controlled comparison of different treatments.

These objectives help to indicate the type and amount of information to be asked. Because of the difficulty in interpreting any apparent causal association, it is undesirable to ask for more information than a necessary minimum. Moreover, if unavailable information is requested, there may be many blanks; or unrecorded information may be entered in our charts from memory.

Note.—To reduce the risk mentioned in the last sentence of the above statement of aims, the questionnaire charts prepared by the committee contained after each question the letters "NR," to be circled if there was no record of the information requested.

COMPLEXITY OF APPARENTLY SIMPLE SURVEYS

Although clinicians may readily recognize the difficulty of such surveys as have been discussed here, they often consider it a simple matter to find the association between a particular disease and some biochemical, physiological, structural or environmental feature. They may wish, for example, to find whether the average blood cholesterol level is unusually high in patients with a certain disease, or whether a disease occurs more frequently in persons of a certain blood group than in persons of other groups, or to ascertain the age distribution of a certain heart lesion found in autopsies.

For such a purpose the investigator may plan to compare the data from the subjects in question with data from 1 or more of a variety of sources, such as: patients (or autopsy specimens) in the same hospital with other diseases; hospital staff or medical students; published data (on healthy or diseased subjects) from the same geographical region or elsewhere. The observations and comparisons are often not difficult to make, and the contrasted groups may show a highly significant difference. The real difficulty is in the interpretation, even after allowance has been made for differences in technique, the race, sex and age of subjects, the occurrence of the same disease in more than 1 member of the same family, and other known relevant features. The resulting figures usually seem so simple that it is hard for anyone who has not studied Berkson's fallacy to realize that the simplest figures may in effect tell nothing because they are open to so many possible explanations.

This does not imply that all such researches should be discouraged, for they may provide helpful hints for investigations by more reliable methods. To anyone contemplating a study of this kind the following procedure is suggested:

1. Before starting to collect data, visualize the proper method

of obtaining a sample for comparison with the subjects under study—a sample that would permit an unequivocal interpretation. This would be a strictly random sample of the population (without the disease) that would go to the particular hospital if they had that disease.

2. Try to think of all features in which the sample that you propose to use instead of the proper sample differs, or might differ, from the proper sample.

3. As a result of (2), make a list of all the assumptions (to be published in your report) that must be made in using your actual sample instead of the proper sample.

4. State the evidence that seems to indicate that the assumptions in (3) may be fairly safe.

5. Discuss the project with a statistician who is familiar with such problems.

REFERENCES

1. Belk, W. P., and Sunderman, F. W.: A survey of the accuracy of chemical analyses in clinical laboratories, Am. J. Clin. Path. 17: 853, 1947.
2. Berkson, J.: Limitations of the application of fourfold table analysis to hospital data, Biometrics Bull. 2(3): 47, 1946.
3. Cochran, W. G.: Modern methods in the sampling of human populations: General principles in the selection of a sample, Am. J. Pub. Health 41: 647, 1951.
4. Cornfield, J.: Modern methods in the sampling of human populations: The determination of sample size, Am. J. Pub. Health 41: 654, 1951.
5. Hansen, M. H., and Hurwitz, W. N.: Modern methods in the sampling of human populations: Some methods of area sampling in a local community, Am. J. Pub. Health 41: 662, 1951.
6. Mainland, D.: *Elementary Medical Statistics: The Principles of Quantitative Medicine* (Philadelphia: W. B. Saunders Company, 1952).
7. Mainland, D.: The risk of fallacious conclusions from autopsy data on the incidence of diseases with applications to heart disease, Am. Heart J. 45: 644, 1953.
8. Milbank Memorial Fund: *Research in Public Health*, Papers Presented at the 1951 Annual Conference of the Milbank Memorial Fund, New York (New York: Milbank Memorial Fund, 1952).
9. Ogilvie, H.: The use of experience, Brit. M. J. 2: 663, 1949.

SOME UNDESIRABLE EFFECTS OF LABORATORY TRADITION

DONALD MAINLAND

One of the impediments to the introduction of modern statistics into medicine has been the influence of academic chemistry and physics on which much medical research was founded. Workers in those 2 sciences were familiar with experiments in which, after standardization of technique and experimental material, the residual variation, due mostly to observational error, was small compared with the effects to be measured. It could often be assessed by inspection of the data or by a simple calculation. Such assessments, although statistical in nature, were quite unsuitable to cope with biological variation, but many laboratory workers failed to appreciate this. Moreover, the habit of eliminating variables in an academic laboratory does not equip an investigator for research in an applied science like medicine, where the information, to be of practical value, must be applicable to situations in which many factors, often interacting with each other, are present. Indeed, an engineer is more likely to appreciate the methods necessary in medical research than is an academic physicist.

The attitude of academic laboratory workers is changing, and even in "pure" chemistry and physics the methods developed in biological statistics are now being adopted, both in design of experiments and in analysis of results (9); but there remain in medical research several features that are, to some extent at least, relics of an earlier tradition. Five of these will be discussed under the headings: (1) percentage change; (2) percentage frequencies; (3) the variable error of an observational method; (4) graphs, correlation and curve fitting; (5) rejection of outlying observations.

1. PERCENTAGE CHANGE

When a value has risen or fallen between 1 observation time and the next (e.g., an increase in urinary steroid level after the administration of a steroid hormone), it is not uncommon for the change to be expressed as a percentage of the initial value. An observer is, of course, at liberty to express a measurement in any form that he chooses, but he should know: (1) the reason, if any, why the particular form is adopted; (2) the possible effect of the

expression chosen upon comparisons that he may make between 1 set of observations and another.

Often the percentage expression seems to be used automatically, as a mere habit; but if asked for the reason the observer may say that it is desirable to standardize the change (e.g., in urinary steroid level) found in various individuals under the same experimental conditions—to bring them all to the same scale by making allowance for the initial differences in level between the individuals, as one would feel necessary in comparing the growth of a mouse with that of a child. This implies, however, that the change varies in proportion to the initial level—the higher the level the greater the change—and often the readings themselves show no evidence of this.

In 10 patients at the 4th hour of a blood sugar experiment there were administered 100 mg of a certain pituitary hormone and 0.1 units of insulin per kg of body weight, and the following levels (mg/100 ml) were found:

Patient	4 Hr	4.5 Hr	Diff.	Diff. as % of 4-Hr Level
1	72	61	−11	−15.3
2	89	33	−56	−62.9
3	93	35	−58	−62.4
4	79	33	−46	−58.2
5	90	29	−61	−67.8
6	89	63	−26	−29.2
7	85	52	−33	−38.8
8	86	39	−47	−54.7
9	86	29	−57	−66.3
10	94	71	−23	−24.5

The lowest initial reading (72 mg) was followed by the least change (11 mg), but inspection of the data does not suggest that the higher the initial level the greater is the difference between the initial and final levels, i.e., with the 1 exception there is no indication of a relationship between the initial reading and the change.

For greater clarity it is often useful to plot such readings as a dot diagram (scatter diagram) with the initial level as X and the difference as Y; and in this instance it will be seen that, except for the outlying item ($X = 72$, $Y = 11$), there is no suggestion of a trend such as would be represented by an upward or downward sloping regression line. Instead of finding the actual regression line it is convenient to find the coefficient of correlation between X and Y, and test its significance, for this is the same as testing the significance of the slope of the regression line. The coefficient of correlation in this case was 0.43, and with 10 pairs of readings this is far from significant at the 5% level, for P, the probability of chance occurrence, is between 0.3 and 0.2. There is, then, no evidence of a relationship between the magnitude of the initial level and the magnitude of the change in level.

To indicate in a general way the magnitude of a change it may be help-

ful, after all the necessary tests have been applied, to express it as a percentage of the initial value, thus: Mean at 4 hr = 86.3 mg; mean at 4.5 hr = 44.5 mg; difference = 41.8 mg, i.e., 48.4% of the initial level. *Note:* Because this is merely a kind of graphic device, not aiming at precision, it is unnecessary to find the percentage change for each individual and then calculate the mean of these percentages. (In the present case there would be little difference from the above figure—48.0% instead of 48.4%.)

The expression of a change as a percentage does not usually involve simply 1 set of data as above, but is carried through the whole analysis. In the experiment from which the data on 10 patients were cited was another sample of patients to whom a different dose of the pituitary hormone had been administered. The differences in blood sugar level were expressed as percentages of the initial levels, and the investigator proposed to compare (by the *t* test) the mean percentage changes from the 2 samples, using percentages in order to compensate for any difference that might exist between the initial levels in the 2 groups. Three points should be noted regarding such a procedure:

1. If there is no significant difference between the initial readings in the two sets of patients (and a difference significant at the 5% level would occur in only 5% of experiments with proper random allocation of treatments), there is no reason to attribute a significant difference in the change in readings to the initial differences. Unless, therefore, there is a special reason for studying the relationship between initial level and change in level, it is sufficient to make a comparison of mean differences in absolute units (mg), not making allowance for initial levels.

2. Even when it is desirable to make allowance for the relationship between initial level and the change in level, the percentage form is often not the proper one to express the relationship. It is far safer to base any corrections or adjustments on the relationship that actually holds in the data themselves, in this case any correlation that exists between the change in level and the initial level. This is done by analysis of *covariance*, i.e., variation of 1 variable (e.g., Y = change in level) "along with" another variable (e.g., X = initial level). This method allows for any regression relationship between the 2 variables—the slope of the line representing the relationship of the 2 variables. In the present example it would automatically apply an adjustment based on the difference between the initial levels of blood sugar. The method is not very difficult for those who have done some analysis of variance. Examples are found in textbooks of biological statistics (8, 7).

3. Another undesirable feature of the "percentage change" expression is that it produces quantities that may be quite unsuitable for treatment by methods such as the *t* test, which are derived from the normal (Gaussian) curve.

The nitrogen balance was found in 8 patients before and after the administration of androgenic hormones:

Before (B)	After (A)	Difference (D)	D as % of B
+1.08	+2.80	+1.72	+159.3
−4.38	−0.50	+3.88	+ 88.6
−0.30	+2.65	+2.95	+983.3
−0.92	−1.30	−0.38	− 41.3
+1.30	+1.22	−0.08	− 6.2
+2.65	+6.05	+3.40	+128.3
+2.52	+2.80	+0.28	+ 11.1
−0.88	−0.62	+0.26	+ 29.5
		Mean = +1.504	Mean = +169.1

The investigator proposed to test the mean percentage difference. Although 8 items are insufficient to indicate clearly the shape of the frequency distribution of percentages, extremely high values are obviously possible, because if the initial reading is very near zero a slight absolute change can register as a very high percentage change. The *t* test showed that the mean absolute difference (+1.504) was significant (P less than 0.05), but the mean percentage difference was far from significant (P approximately 0.2), and this was obviously due to the great variation among the percentages. The proper figures to use are the absolute differences, and if afterward it is desired to explore the percentage relationship (e.g., to find confidence limits for the mean percentage difference), a special method must be used (3).

2. PERCENTAGE FREQUENCIES

When observations are made not by measurement but by enumeration (as in blood counts or numbers of patients showing improvement under a certain treatment), percentage expressions are even more treacherous than in measurement data, and an academic laboratory worker's experience, so frequently involving measurements, does little to fit him for handling percentage frequencies. One drop of a 5% solution of a certain chemical is so different from one drop of a 30% solution that it is difficult to believe that a sample of 20 mice, of which 5% have tumors, does not differ significantly in tumor incidence from another sample of 20 mice containing 30% with tumors.

Even when a proper test has shown a significant difference in frequencies, percentage expressions can be confusing. If, treated by 1 method, 20 patients of 100 die, whereas treated by another method only 5 patients of 100 die, the case fatality rate is "300 per cent greater" in the first case than in the second; but the same

fact can be expressed by saying that the rate has been "reduced by 75%" or that the recovery rate has been "increased by 15/80, i.e., 18.75%." The choice of expression usually depends on whether a writer wishes to minimize or emphasize a difference. The most objective statement is: Difference = 20 − 5 (or 95 − 80) = 15% (or 15 percentage points).

3. THE VARIABLE ERROR OF AN OBSERVATIONAL METHOD

In biochemical and physiological experiments 2 or more readings are frequently taken in order to provide an estimate of the variable (random) error, as an indication of the precision, consistency, or reproducibility of the method. This kind of observation should be made, much more frequently than it is, in nonmetrical assessments, where, for example, a degree of hyperemia is expressed as +, ++, or +++, and even where the verdict is of the "yes or no" type, as in many radiographic assessments. Great inconsistency in radiographic diagnosis of pulmonary tuberculosis was revealed in this way (1). Besides differences among the 5 radiologists who participated, there were frequent reversals of both positive and negative verdicts when the same 5 re-examined the films without knowing their previous verdicts.

The determination of observational variation (in the same observer and between different observers) seldom necessitates the setting up of tests separate from the main experiment. Indeed, such a separate test is generally undesirable except as a periodic check of routine analysis made for diagnostic purposes. In research the object is to see how far the observational error will account for the differences which the experiment is designed to measure, and therefore the assessment of error should be incorporated in the experiment itself. With careful planning this can usually be done, and it will at the same time increase the precision of the experiment. There are, however, several pitfalls and sources of misunderstanding which can be considered under the headings: (1) independence of readings; (2) completeness of replication; (3) estimation of variable error; (4) balance in experimental precision.

Independence of readings.—The members of each pair, trio or other set of readings (duplicates, triplicates or other replicates) should all be independent of each other. Sometimes pairs of readings, for example in blood chemistry, handed to a statistician for estimation of the variable error do not meet this requirement, for application of the "sign test" (p. 209) shows that in a significant majority the 1st reading is higher (or lower) than the 2d. This can arise in various ways. There may have been an actual change in the blood between the 2 readings, or the observer may have used

the 1st reading as a guide in making the 2d, as is often done in a titration involving a color change. Whatever the cause, a systematic difference is present, i.e., the 2 readings are not strict replicates—independent readings suitable for the estimation of variable error. Indeed, if the quantity that is being measured changes between the 2 readings, no estimation of that error is possible for the method employed.

An analogous lack of independence was seen when an observer demonstrated how he assessed (on a +, ++, +++ scale) the degree of hyperemia resulting from different exposures to a radioactive plaque. From the sequence of the areas on the patient's skin, he knew the gradation of doses (exposure times) even if he did not remember the actual amounts. Any attempts to test the reproducibility of his assessments by repetition of the process, or to compare his assessments with those of another observer, could not use the assessments of successive areas as if they were independent.

Completeness of replication.—Lack of independence occurs also when the so-called replication does not repeat all the steps of a procedure. When a series of experiments in physiology was being designed to employ mixtures of radioactive sodium and radioactive potassium, preliminary tests were run on solutions of salts of these 2 substances. On receipt of a sample of the sodium salt and of the potassium salt, solutions were made and mixed in a certain ratio, and duplicate determinations on aliquots of the mixture were made by Geiger counter. The procedure was repeated when other samples of the same 2 salts were received from the atomic energy laboratory, but it was found that the results from the various samples were too discordant to be explained by the variable error as estimated from the duplicate determinations. It was then found that the duplication had been only partial, for it had not included the dissolving, dilution, and mixing of the salts, and thus a considerable part of the variable error had not been measured.

Estimation of variable error.—The method of estimating variable or random error should enable the investigator to estimate what allowance to make for it in analyzing the results of an experiment. In practice, a variety of methods, derived originally from the practice of physicists or chemists, are brought to the statistical laboratory. Such are: the mean deviation; the maximum difference found in a set of duplicates or other replicates; ± half the difference between the extremes in a set of replicate readings. Often these or other estimates are expressed as "percentage errors," which is open to objection on the grounds already discussed. Except for this, there would be no serious objection to any method, provided

the investigator (1) used it consistently, (2) knew what his estimates implied regarding the allowance to be made for variable error, and (3) was able to incorporate his estimate in the complete analysis of his experiment. These conditions are, however, seldom fulfilled by estimates such as are mentioned above.

The method that fulfills the conditions most closely and easily is the *standard deviation* (*standard error*) *method*, which has been long used in biometry, and which, with duplicate readings, consists of 4 simple steps: (1) square the difference between the duplicate readings; (2) summate these squares; (3) divide the sum by twice the number of pairs; (4) find the square root of the quotient in (3). This square root is the standard deviation of individual readings, i.e., it expresses the variation among the readings, and tables of the normal (Gaussian) distribution (or its derivative, the t distribution) enable one to use it in estimating the allowance to make for random observational error on a definite probability basis. (If more than 2 readings are taken, a slight modification of the technique is all that is required.)

The standard deviation method is the one that fits best into the complete analysis of the experimental results. Indeed, it is automatically part of such an analysis, which can contain, for example, the following classes of variation: between treatments, between animals on the same treatment, between observers, and between replicate observations by the same observer on the same animal.

Balance in experimental precision.—Great effort at high precision in 1 part of an experiment may be wasted if the precision in another part is low. In an investigation of salt exchanges between the intestine and blood in animals, chemical analyses were required, and the experimenter said that he felt that the variable error in these analyses should be kept low, to about "2%." Apparently he would have been ashamed to publish results containing a greater error than this, and he asked how many repetitions of each analysis he should make in order to attain this precision.

It is, of course, desirable to use chemical methods that are dependable in the sense of not introducing every now and then a grossly discordant estimate; but the actual precision to be demanded depends on the particular experiment. In the investigation of salt exchange, the variation in the results between experiments would depend partly on the precision of the chemical analyses, but also very largely on the variation between animals (or between different experiments on the same animal). Even at an early stage of an investigation it often becomes obvious that the biological variation will be much greater than the variation in chemical analysis, and therefore increase in the chemical precision,

either by improved methods or by more determinations on the same specimen of fluid, would have little effect in reducing the variation in the experiment as a whole.

When an experiment is done in such a way that analysis of variance can be applied to the data, there is a method for estimating the relative weights of the different sources of variation—the analysis of *components of variance* (8). If, for example, a certain volume of fluid is taken once from each animal in an experiment and aliquots of this are analyzed chemically, all by the same method, the analysis of the data can take the form: variation between experimental treatments, variation between animals on the same treatment, variation between chemical determinations on the same animal. By a further step it is possible to estimate the relative advantages of (1) increasing the number of readings per animal and (2) increasing the number of animals, and this may show that a tenfold, or even a hundred-fold, increase in the number of chemical analyses would do hardly anything to reduce the variation between animals.

Another difficulty regarding the variation between readings is illustrated by the question sometimes asked by experimenters who say: "The statistical analysis has shown a significant difference between the effects of the treatments, but the difference is less than the error (5%) in the chemical analyses that contributed to the data. How is that possible?" The answer is that the error in the chemical analyses represents differences between individual determinations or readings, whereas the analysis of the data compares the mean effects of the different treatments. In comparing mean statures of males and females we do not use directly the variation between individuals, but estimate from them the variation between mean statures. Analysis of experimental results does essentially the same thing, allowing to each source of variation (between animals, between chemical determinations, and so on) its appropriate weight.

4. GRAPHS, CORRELATION AND CURVE FITTING

Graphs.—Graphs of data are so useful that they are often misused; and the warnings of 2 experienced statisticians deserve frequent quotation:

"Diagrams prove nothing, but bring outstanding features readily to the eye; they are therefore no substitute for such critical tests as may be applied to the data, but are valuable in suggesting such tests" (2).

"Graphs should always be regarded as subsidiary aids to the intelligence and not as the evidence of associations or trends" (5).

An even more obvious principle is that graphs should clarify a relationship, but many of those that arrive at the statistical laboratory, sometimes without accompanying tabular data, do the reverse. They may contain 10 or a dozen criss-crossing lines each in a different color, or with individual subjects distinguished by symbols such as large dots and small dots, circles, squares, and triangles. Investigators often seem to be unaware of the tricks that the eye can play in creating a pattern or in masking a real trend.

In order to improve the quality of one's graphs, the learning of rules is not nearly so useful as the copying of good models and the inspection of one's efforts by another worker with only the text of the accompanying report as guide to interpretation of the graph.

Correlation.—Laboratory and clinical workers are often misled when they use correlation techniques. Some of the chief difficulties are discussed elsewhere (6), and it is recommended that the emphasis should be placed, not on correlation, but on regression. The full, and complex, implications of correlation are not applicable to many of the phenomena met in medical research, whereas regression expresses directly what the investigator wishes to discover—a trend represented by a line in a graph, e.g., a change of pulse rate with the passage of time. The regression coefficient expresses the slope or gradient of this line. In many medical research problems, therefore, the 2 chief (or only) uses of the correlation coefficient are:

1. To provide a simple test of the significance of the regression coefficient. This is particularly useful when the investigator wishes to know whether the data provide evidence that a slope is not adequately accounted for by chance, but does not need to know what the gradient is.
2. To indicate the scatter of the individual measurements around the regression line derived from them.

Curve fitting.—In chemistry and physics there often is some theoretical reason for believing that a particular kind of curve, representing a certain type of reaction or relationship, will be appropriate to a given set of measurements; or, quite commonly, the question arises: Which of these 2 types of curve, representing different types of reaction or relationship, fits the data better?

To answer such a question, the curves, with numerical values calculated from the data, are inserted in the graph along with the actual measurements, and it may then be proclaimed that 1 curve gives an obviously better fit than the other. An investigator who is well acquainted with the small observational error of his experi-

mental techniques may be quite justified in this assertion; but the method is often applied in physiology and biochemistry where the variation, biological and observational, is considerable and there is no precise knowledge of its magnitude. Appreciating this, the investigator may apply a numerical test. For example, he may, for each curve, find the sum of squares of the differences between the individual measurements and the curve. If the deviations from curve A are less than from curve B, he may take it as sufficient proof that curve A is the correct one, or is more likely to be the correct one. This is, however, quite inadequate proof. The question to be asked is: Are the deviations *significantly* less from curve A than from curve B? To answer this question, there has been developed a routine method (2, 8, 7) of testing successively a series of curves when they belong to the polynomial series (Y expressed in terms of X, X^2, X^3, etc.). Comparing other types of curve, e.g., a logarithmic with a polynomial, presents a little more difficulty, but methods are available (8).

Some medical and biological investigators, with chemical or physical background, object to the fitting of polynomial curves, such as the parabola ($Y = k + b_1X + b_2X^2$), to biological data because they seldom represent the true relationship or law connecting X and Y. The validity of the criticism depends on the purpose of the curve fitting. When we fit a straight line (the simplest form of curve fitting), we commonly wish, not to demonstrate that the relationship between X and Y is strictly rectilinear, but merely to find out whether there is, in general, an upward (or downward) trend in Y as X increases. We are testing the impression created by the figures, or by the dots in a graph. Similarly, we often wish to test the impression that the relationship between X and Y is curved, e.g., an upward slope of gradually diminishing steepness or an upward slope followed by a downward slope. Fitting a parabola enables one to test this impression, without the implication that it represents exactly the law relating Y to X.

Regarding curve fitting in general, the following quotation from Snedecor (8) is apposite:

"A stupendous amount of time has been wasted in ill-advised curve fitting. Only when the end in view is clear should the task be undertaken. Often a graph of the data points is sufficient. Represent them by small circles or heavy dots. If desired, they may be connected by light line segments. Avoid drawing 'eye-fitted' curves. They are highly subjective and are apt to be misleading to both the perpetrator and the victim. Interpolation with these links stands a better chance of being good than does estimating by means of even the most artistic curves."

Finally, it should be noted that in order to test for curvilinearity it is not always necessary to fit a curve. With suitable data, 1 or both of 2 methods are available:

1. If for each X value there are 2 or more Y values (not necessarily equal numbers for each X), there is, as part of the analysis of variance, a fairly simple test for departure from linearity (2, 8).

2. If the curve under consideration is such that by transformation of variables (e.g., replacement of the actual readings by their logarithms) a straight line will be produced, the simple method of straight line regression can be applied to the transformed variables (7).

5. REJECTION OF OUTLYING OBSERVATIONS

The problem of when to reject from a series of measurements one or more that seem out of line with the rest is difficult and controversial. A physicist or chemist, well acquainted with the random error of his methods, may be justified in rejecting a reading because its magnitude shows that it is very probably due to a gross mistake. Sometimes, even in physics, such judgments have subsequently been shown to be wrong; but the serious danger arises when these practices are extended to physiological and biochemical data which (1) are obtained by techniques with large and inadequately known random error, and (2) contain a large biological variation. When a biochemist proposed to reject a vitamin A determination from the liver of 1 fish out of a dozen of the same species, his proposal arose not from any knowledge of a fault in the technique, the technician, or the fish, but apparently reflected simply his chemical training and lack of acquaintance with the skewness of biological distributions.

The general rule should be not to reject observations without adequate knowledge of observational and biological variation. This rule does not imply an uncritical acceptance of aberrant data. If the technique is suspected it should be investigated and, if possible, made more reliable; and if the unusual values are due to biological variation they suggest either an unsatisfactory selection (or standardization) of material or a marked skewness of the frequency distribution of the quantities measured. In this connection it should be noted that various rules proposed for the rejection of observations according to the variation found among the other readings are based on the assumption of a normal (Gaussian) distribution of readings, and they are therefore inapplicable to many biological data. (A simple rule, such as the selection of the "best two out of three" readings, can distort the data greatly).

If, in the absence of a clearcut reason for rejecting an observation, it is included in the analysis when it ought to have been rejected, the tendency will often be for it to increase the variation against which the treatment differences are tested. Therefore the treatment difference will tend to be pronounced nonsignificant when otherwise it might have been called significant. Generally this is a safer error than one in the opposite direction—a verdict of significance due to the improper exclusion of an observation that ought to have been included.

REFERENCES

1. Birkelo, C. C., *et al.:* Tuberculosis case finding: A comparison of the effectiveness of various roentgenographic and photofluorographic methods, J.A.M.A. 133: 359, 1947.
2. Fisher, R. A.: *Statistical Methods for Research Workers* (Edinburgh and London: Oliver & Boyd, Ltd.; New York: Hafner Publishing Company, 1948).
3. Fisher, R. A.: *The Design of Experiments* (Edinburgh and London: Oliver & Boyd, Ltd.; New York: Hafner Publishing Company, 1947).
4. Fisher, R. A., and Yates, F.: *Statistical Tables for Biological, Agricultural and Medical Research* (Edinburgh and London: Oliver & Boyd, Ltd.; New York: Hafner Publishing Company, 1948).
5. Hill, A. B.: *Principles of Medical Statistics* (London: The Lancet, 1945).
6. Mainland, D.: *Elementary Medical Statistics: The Principles of Quantitative Medicine* (Philadelphia: W. B. Saunders Company, 1952).
7. Mather, K.: *Statistical Analysis in Biology* (London: Methuen & Co., Ltd.; New York: Interscience Publishers, Inc., 1946).
8. Snedecor, G. W.: *Statistical Methods Applied to Experiments in Agriculture and Biology* (Ames, Ia.: Iowa State College Press, 1946).
9. Youden, W. J.: *Statistical Methods for Chemists* (New York: John Wiley & Sons, Inc., 1951).

INDEPENDENT INDIVIDUALS

DONALD MAINLAND *and* LEE HERRERA, *New York University*

IN A DISK sampling experiment (p. 127) each disk is an independent individual or unit, because although each undoubtedly shares part of its "career" with groups of other disks, if the mixing is thorough these links are continually being broken. It is such independence that qualifies an item in a sample to be treated as an individual when, for instance, the means of 2 samples are compared by the *t* test.

In translating the sampling experiment into clinical and laboratory terms each disk was, for simplicity of discussion, equated to a patient or an animal; but it is important to realize that in any particular experiment an individual in the colloquial sense may not be an independent unit.

ANIMALS AND CAGES

Animal experiments illustrate very well the risk of confusion regarding individuals; and several forms of such experiments are worth while examining.

One treatment per cage.—In the simplest case, 4 animals in 1 cage are subjected to treatment *V*, 4 animals in another cage to treatment *W*. The mean responses are compared by the *t* test, the standard error being derived from the variation among the animals in each cage. The implicit assumption here is that the only causes of difference between the 2 sets of animals are (1) variation between animals in each cage, and (2) the difference between the treatments *V* and *W*.

The validity of this assumption, however, is not assured by the experimental method. The animals may, indeed, have been independent at the outset, but even the proper method of insuring this (random allocation of the 8 animals, 4 to a cage) does not insure that the individuals in a cage will remain independent except for the treatment that they receive in common. The investigator may use uniform cages and may try to secure uniform light and ventilation (or may randomly allocate the cages on the animal house shelves). He may find no evidence that 1 animal has affected others in the same cage, e.g., by spread of infection, or by domination at the food supply. But all experimental precautions and ancillary evidence, aimed at supporting the assumption of independence of

individual animals, are not equivalent to an experimental design that insures the maintenance of this independence.

All factors, apart from the treatments V and W, that may affect in common all animals in a given cage can be called "cage effects," and the experiment just described has *confounded* cage effects and treatment effects; i.e., it has confused or mixed the 2 (real or potential) sources of variation in such a way that they cannot be distinguished. Expressing this otherwise, we can say that there was only 1 "individual" or independent unit (1 cage of animals) on each treatment. There were no *replicates*, i.e., independent individuals subjected to the same treatment.

More than one cage per treatment.—Usually, of course, more than 1 cage of animals receives the same treatment, say 5 cages (each containing 4 animals) on treatment V, and 5 cages on treatment W. In analyzing the data from such an experiment, however, experimenters often disregard possible cage effects. With 40 animals (4 per cage) they would compare the mean response of the V-treated and the W-treated animals as if there were 20 independent individuals in each sample. The t test, so applied, asks the question: Is the difference between the V-treated group and the W-treated group significantly greater than the variation between independent animals treated alike? The experiment really contains, however, only 5 independent individuals (cages) in each sample, and the question that it can answer is therefore: Is the difference between the 2 treatment groups significantly greater than the variation between independent lots (cages) of 4 animals treated alike?

The simplest way to eliminate cage effects, in order to test treatment differences against animal differences, is to allocate all the treatments to all the cages in a *randomized block* design, which was developed in agriculture. In animal experiments the cages are blocks, corresponding to areas of land, and the animals correspond to plots of land within the blocks. In each cage the treatments must be distributed at random. Further details of planning, e.g., the placing of similar animals (litter mates or animals of equal weight) together in a cage, and the question of random allocation of cages on the animal house shelves, must be decided in each experiment.

The foregoing scheme enables one to test as many treatments as there are animals per cage, but it does not allow one to test the possibility that the difference between treatments may vary from 1 cage to another, i.e., a treatment-cage *interaction*. To be able to test for this interaction it is necessary to apply the treatments to more than 1 animal per cage, e.g., by random allocation of treatment V and treatment W to 2 animals in each 4-animal cage.

Single-animal vs. multi-animal cages.—Although the designs just mentioned, or more complicated ones, permit the separation of the sources of variation, the housing of more than 1 animal in a cage is open to several objections:

1. There is, in general, greater risk of spread of disease among cage mates than among animals in separate cages.

2. Sickness or death of animals can easily unbalance an experiment that uses multi-animal cages. To compensate for 2 or 3 such incidents, estimates can often be made in such a way as to avoid bias; but when catastrophes are more numerous the whole experiment may be ruined.

3. When all animals in the same cage do not receive the same treatment there is a risk of exaggeration of treatment effects. If, for example, 1 treatment makes an animal feel temporarily sick it may, during that period, be pushed from the food dish by its healthier fellows and become undernourished. The resulting weight loss may be attributed directly to the treatment, and in addition the state of the animal may produce a response to treatment (or in some cases a lack of response) that would not be found in an animal that had ample opportunity of obtaining nourishment.

The provision of a cage for each animal adds, of course, to the cost and the labor; but against these factors should be weighed the various disadvantages of multi-animal cages.

RESEARCH ON HUMAN BEINGS

The foregoing discussion of animal experiments should enable the reader to detect many analogous instances in research on human beings. If in 1 school all children receive a dietary supplement for contrast with children in another school where this supplement is not given, there are only 2 independent "individuals" (schools) in the experiment, and, since each receives a different treatment, the treatment effect cannot be tested. Similarly, if 1 clinical treatment is applied in 1 hospital ward and another treatment in another, however many patients are in each ward there are strictly only 2 independent individuals in the experiment. The investigators may demonstrate great similarity between the 2 groups of patients, between the routines in the 2 wards, and so on; but any conclusion regarding the difference (or lack of difference) in the effects of the 2 treatments must ultimately rest on the unprovable assumption that the wards do not differ significantly in degree or kind, with respect to any feature (physical or psychological) except the 2 treatments under test.

PAIRED READINGS

The concept of independent individuals may help to remove a difficulty that some research workers still find in handling paired readings. Having taken "before and after" readings on a number of animals or patients, such as blood pressure before and after an injection, or weight before and after a period of experimental feeding, they sometimes ask whether they ought to compare the mean "before" with the mean "after," or find the difference for each animal or person and test the significance of the mean of these differences.

Many workers know that the latter method is correct when each pair of readings is made on the same subject, but they may express doubt when the members of the pairs are different subjects, e.g., litter mates. In this case, let it be supposed that treatments V and W have been allocated strictly at random within each pair. That is, the design of the experiment has been based on the knowledge or hope that in each pair, except for the possible effect of V or W, the readings would be more alike than between unmatched subjects. Therefore the V-treated animals and the W-treated animals cannot be considered as independent random samples. The independent items are the differences between the readings, $V-W$, in each pair; and therefore the proper test, as when both readings are made on the same subject, is a test of the mean of the differences.

The foregoing statement may indicate why warnings are given to experimenters against "artificial" or "injudicious" pairing. If in the V vs. W experiment pairing was not used, but the treatments were randomly allocated to 20 animals (10 animals on each treatment), the difference $V-W$ would be tested against the individual variation found in 2 lots of 10 animals. When pairing is used, the difference $V-W$ is measured against the variation among only 10 individual items (differences). Unless, therefore, the pairing is based on something more than mere hope that the members of a pair would respond similarly to similar treatment, it is not an advisable technique.

MULTIPLE DETERMINATIONS FROM THE SAME SUBJECT

If anyone wished to estimate the mean and individual variation in stature of boys of a certain race and age, he would not do so by measuring 2 boys, each 100 times, and using the observational variation (although it is often considerable) as a measure of variation between boys. Nevertheless, analogous procedures are not uncommonly met in data brought to the statistical laboratory.

Let us suppose that the concentration of creatinine in the urine is estimated on 9 animals of a certain species, and that all the ob-

servations on any 1 animal at different times are strict replicates, obtained without influence of changed conditions. The numbers of observations vary from 2 to 10 per animal. The physiologist wishes to use all 50 readings in order to express the concentration in the form of a mean and a standard deviation. For this purpose 2 methods are commonly applied to such data, and, as always in evaluating a method, one should specify the population that the data represent.

1. Each observation is treated as if it were a boy's stature measurement, each boy having been measured only once; and the mean, standard deviation of series, and standard deviation of mean are calculated. The creatinine readings would thus correspond to a sample of 50 boys' statures. The population represented is, however, a mixture of individual animals and readings from the same animal; and the method just described would be permissible only if it were known that the readings from any 1 animal would be equivalent to single readings, each from a different animal. Even if the number of readings from each animal were the same, the objection would remain. Nor could it be removed by showing that there was, among the animals in the experiment, no significant difference between the mean readings from the various animals. To defend the pooling of readings and animals by this verdict of nonsignificance would be analogous to doing the same thing with the sample of 2 boys' statures (with 100 readings each) if their mean statures were so alike that the difference was found to be nonsignificant when tested against observational error.

2. The mean for each animal is found. The creatinine sample would thus provide 9 mean values, which would be treated as if they were statures of 9 boys, to give a mean and a standard deviation for the series. This standard deviation gives indeed an expression of variation between animals, but when the numbers of readings per animal are unequal the method is open to objection, as can be seen by exaggeration. Let it be imagined that the mean for 1 animal was based on such a large number of replicate readings that it would be essentially the "true" value for that animal (as far as determinable by the method of chemical analysis used); i.e., the effect of further readings obtained from that animal under the same conditions would be negligible. At the other extreme, if the means from some of the animals were derived from only 2 readings each, the variation between these means would be due in considerable measure to the variation between the readings, and further readings on the same animals would be likely to change their relationships to each other.

Therefore the population to which one can argue from data con-

taining unequal numbers of observations per animal is a population in which the same proportion of animals would provide the same number of readings per animal. For example, if 10% of the animals in the sample provide 2 readings each and 15% provide 6 readings each, these proportions would have to hold in the population—a very inconvenient restriction and undesirable in most experiments.

The best method of collecting and analyzing such data is to take the same number of readings, 1 or more, on each animal or other subject, and find the mean for each subject. If there are, say, 2 readings per subject, the general mean will be an estimate of the population mean that would be approached by taking more subjects, and the standard deviation of the series will be an estimate of the intersubject variation between such 2-reading means. Further, the standard deviation of the general mean can be estimated in the usual way by dividing the intersubject standard deviation by the square root of the number of subjects.

Although the foregoing discussion may appear hardly necessary to anatomists, physiologists and biochemists who are investigating such features as dimensions, reactions, or excretions, that are obviously an intimate part of the animal or human subject examined, it seems to need special emphasis in cases where the connection is less intimate, for example in parasitology. A batch of eggs of a certain kind of worm was divided into 4 equal portions and 1 portion was injected into each of 4 monkeys, 2 of which were subjected to treatment *A* and 2 to treatment *B*. After the worms had developed, 25 of them from each monkey were measured. It was proposed that the mean sizes for the *A*-treated and the *B*-treated be compared by the *t* test as if there were 2 samples each containing 50 independent individuals.

There is, however, no reason to suppose that worm length is entirely unaffected by the host; and even if another experiment, on a large number of monkeys all treated alike, had shown that the difference in worm length between monkeys was narrow, it would be unsafe to assume that this would hold for the particular experiment under discussion. In that experiment the independent "individuals" were *mean* worm lengths, 2 from the *A*-treated and 2 from the *B*-treated monkeys.

Unequal samples in correlation and regression.—Inequality of sample size provides 1 of numerous pitfalls for laboratory workers and clinicians who use correlation and regression techniques. In the creatinine investigation mentioned earlier, let it be supposed that on the 9 animals in which urinary creatinine was measured (2–10 readings per animal) the concentration of blood plasma creatinine was measured at the same time. From the 50 pairs of

readings (1 from urine and 1 from plasma) the physiologist wishes to find the correlation between the urinary and plasma concentrations and also the regression relationship, i.e., the difference in urinary concentration per unit difference in plasma concentration.

In such cases it is not uncommon for each pair of readings to be taken as an independent observation, like the stature and weight of an individual person; and here again there is confusion of intra- and intersubject variation. If 1 animal contributes predominantly higher or lower values (from plasma or urine or both) than another animal, the results (correlation and regression coefficients) will differ greatly according to the relative sizes of the contributions (numbers of pairs of readings) from each animal.

The question of relationships should be separated into 2 parts:

1. What is the relationship between creatinine concentration in plasma and urine within each individual animal?

2. When the plasma concentration and the urinary concentration are found for each animal, what is the relationship, from animal to animal, between these 2 concentrations?

To answer the 2d question, the number of pairs of readings from each animal should be the same, 1 or more; and if more than 1 reading is taken, the means for each animal are the values to be used in the analysis. If an answer to the 1st question is desired, at least 3 pairs of readings from each animal are necessary for straight-line regression, and more for curvilinear regression. Although it is not essential that the number of pairs from each animal be the same, equality of numbers greatly facilitates the computation.

CONFIDENCE LIMITS

DONALD MAINLAND

In the estimation of a standard deviation of the mean from a sample of measurements, or of the standard deviation of the difference between 2 means, there has always been implicit the idea of estimating the limits between which the true mean (or true difference) probably lies. That is the conception of confidence limits, the confidence being expressed as a probability that the true value does not lie outside certain estimated limits. The emphasis on significance testing has, however, tended to obscure this idea in the minds of many research workers. It is true that in certain types of statistical work it is necessary to distinguish "tests" from "estimates"; but in some of the routine applications of the simpler methods there is much to be gained from considering significance tests and confidence limits together, and several illustrations of that practice will be given here.

CONFIDENCE LIMITS OF A MEAN DIFFERENCE

Let it be supposed that the mean of a sample of 10 differences (e.g., blood pressure before injection minus blood pressure after injection in each of 10 subjects) is +20 units and the standard deviation of the mean is 5 units. To test the significance of the mean difference we find $t = 20/5 = 4.0$. Entering Fisher and Yates' table (2) of t with $n = 10 - 1 = 9$, we find that any value of t greater than 2.262 has a probability P less than 0.05. The mean difference is therefore significant at the 5% level. Indeed it is even significant at the 1% level, for $t = 4$ is greater than 3.250, which corresponds to $P = 0.01$.

In making the above test we have asked: Is it unlikely that the true mean difference is as low as zero? The answer has been that if the true mean difference were zero, fewer than 5% (even fewer than 1%) of random samples would give such high values of t as did our observed sample. Expressed otherwise, the probability is more than 95% (even more than 99%) against finding such high values as a result of chance alone when the true value is zero.

We can, however, obtain the same answer by estimating confidence limits. For the 95% limits, we take the value of t found under $P = 0.05$ (5%), at $n = 9$ as before, namely 2.262, and write: Observed mean $\pm t \times$ S.D. of mean $= +20 \pm 2.262 \times 5 = +20 \pm 11.310$. This gives +8.690 and +31.310 units as the limits; and

we then say that there is a 95% probability (odds of 19 to 1) against the true mean lying outside that range. This, of course, excludes zero.

For the 99% confidence limits we take the t value under $P = 0.01$ in the same line of the table (3.250), and this gives $+20 \pm 3.250 \times 5$; i.e., +3.750 and +36.250 units. We can, therefore, have confidence indicated by a 99% probability that the true mean does not lie outside this range or *confidence band*. Zero is again excluded, in agreement with the verdict of the significance test.

USEFULNESS OF CONFIDENCE LIMITS

Estimation of confidence limits is not merely another method of applying a test of significance, for the limits show how wide is the band within which may lie the true value, i.e., the value in populations that could be randomly represented by the observed sample. Such information is often very important, as may be illustrated by applying the method to the data (1) that were used in the 1st presentation (in 1908) of what is now called the t test.

On 10 patients, 2 supposedly soporific drugs, dextro- and levorotatory hyoscyamine hydrobromide, were compared for the gain in sleep that they induced. The mean of the 10 differences (levo- minus dextro-) was +1.58 hr; the standard deviation of the mean was 0.389 hr; and t was therefore 4.06. Since the sample size was 10, the required t values are those used above (2.262 and 3.250), and the mean difference is therefore significant at the 1% level.

The confidence limits (95% band) are: $+1.58 \pm 2.262 \times 0.389$; i.e., +0.70 and +2.46 hr, or 42 min and 2 hr 28 min.

This means that, although in these patients the average additional gain in sleep from the levo- isomer (compared with the dextro- isomer) was approximately 1 hr 35 min, we cannot estimate very precisely what would be the average gain in a long series of similar patients under similar conditions. All that we can assert (with 95% probability) is that the average would not likely be less than about $^3/_4$ hr or greater than $2^1/_2$ hr.

In such a situation, and in similar situations in laboratory work, the investigator, desiring higher precision for the mean, would increase his sample; or, desiring less variability, might experiment with another substance. This illustrates the important fact that *something may be very highly significant, i.e., very unlikely to have occurred by chance, but not be very useful in practical application.*

CONFIDENCE LIMITS OF NONSIGNIFICANT VALUES

Valuable information can also be obtained by estimating confidence limits when a value, such as a difference, has been found nonsignificant. In the experiment on soporifics, when the results

after the administration of the dextro- isomer are considered alone, it is found that for the 10 patients the mean additional hours of sleep (as compared, apparently, with a previous observation without soporific) was +0.75 hr. The standard deviation of the mean is 0.5658, and t is therefore 1.326, with probability P between 0.3 and 0.2. The mean is thus far below the 5% significance level, and there is insufficient reason to believe that the compound affected the length of sleep. Nevertheless, a larger series of similar patients under the same conditions might give strong evidence that there was an actual effect—a significant average difference (gain or loss). The confidence limits show the possible magnitude of this difference. The 95% band is: $+0.75 \pm 2.262 \times 0.5658 = +0.75 \pm 1.28$; i.e., −0.53 and +2.03 hours.

Since the lower limit is below zero, we have the equivalent of the significance test; but the confidence limits reveal something more. It might be known that the drug could not possibly cause a loss of sleep, and therefore the negative value could be dismissed; but the upper limit shows that, for all we can tell from the sample of 10, the drug might give, on the average, about 2 hr of additional sleep.

If confidence limits were estimated in all cases where a value had been found nonsignificant, it would probably help to emphasize a lesson that seems hard to learn—that "not significant" means "not proven." An investigator's reaction to a verdict of nonsignificance appears often to depend on his desires in the particular case. Having reached such a verdict from samples of 10 animals or patients under 1 treatment and 10 under another treatment, if he wishes to minimize the possibility of a real difference he will quote the verdict as if it proved that no real difference existed. If, however, he would like to find a real difference he will say that probably the samples were too small to show the difference.

Frequently, of course, one must *act* as if "no significant difference" meant "probably no real difference," when, for instance, a treatment must be used without further experimentation. In the experiment on soporifics the physician would naturally discard the dextro- compound because of the demonstrated superiority of the levo- isomer; but if the latter had some disadvantage, such as undesirable side effects, he might well desire to test the dextro- compound on more patients, in order to establish a narrower confidence band, i. e., a more precise estimate of its possible effect.

BINOMIAL CONFIDENCE LIMITS

Appropriate methods for estimating confidence limits for various metrical quantities, such as means, differences between means, regression coefficients, and correlation coefficients, are to be found in

statistical textbooks, and they are usually easy to apply. For binomial frequencies, such as percentage of deaths or recoveries, it has been customary to use the standard deviation of the binomial distribution, estimated from the observed sample, and treat it by normal curve methods; thus, the 95% limits would be found from the expression: Observed percentage ±twice the standard deviation. It is still sometimes asserted that this approximation is "sufficiently accurate for most purposes," or that an investigator can "supplement it by more accurate methods when necessary." Few investigators, however, know how far the approximation may lead them astray with the small samples and skew distributions that they frequently meet—when they can safely use the method and when they should not use it. As mentioned on page 131, much more accurate approximations have been developed (2), and from these there have been prepared (3, 4) tables and graphs which not only demand little or no calculation by the investigator but obviate the doubt regarding the safety of the simple normal curve approximation.

For the confidence limits of the difference between 2 percentage frequencies, no completely satisfactory simple substitute for the normal curve approximation has yet been developed.

REFERENCES

1. Fisher, R. A.: *Statistical Methods for Research Workers* (1st ed.; Edinburgh and London: Oliver & Boyd, Ltd.; 1925).
2. Fisher, R. A., and Yates, F.: *Statistical Tables for Biological, Agricultural and Medical Research* (Edinburgh and London: Oliver & Boyd, Ltd.; New York: Hafner Publishing Company, 1948).
3. Mainland, D.: *Elementary Medical Statistics: The Principles of Quantitative Medicine* (Philadelphia: W. B. Saunders Company, 1952).
4. Mainland, D.; Herrera, L., and Sutcliffe, M. I.: *Statistical Tables for Use with Binomial Samples in Medicine and Biology*, in preparation.

STANDARDS OF SIGNIFICANCE

DONALD MAINLAND

IN THE DISCUSSION of random sampling experiments (p. 129) it was stated that an investigator, performing such an experiment in order to assess the results of a laboratory or clinical investigation, would set his "standard of rarity." Technically, this is the *significance level;* and it is important to remember that in all investigations this level must be chosen by the investigator. Statistical tables, such as those showing t values and chi-square values, which in effect give the investigator the results of random sampling experiments, offer a number of levels of significance, indicated by different values of the probability P; for example, P = 0.2, 0.1, 0.05, 0.01.

MEANING OF THE PROBABILITY "P"

In words, P can be expressed as *the probability of occurrence if chance (random sampling variation) alone were operating,* or *the probability of chance occurrence.* Professional statisticians might object to such phrases. They might point out that they imply the probability of occurrence of a particular value of t or chi-square, whereas in reality P indicates the probability of values greater than the tabulated value. They might insist also that it should be made clear whether P refers to both tails of a distribution or to 1 tail only. If, however, they have had much experience in helping investigators who merely wish to understand statistical tests sufficiently to use them safely, they will probably doubt the wisdom of adding obstacles to this understanding. In the beginning, at all events, for the proper use of the tests it is far more important to emphasize 2 points:

1. The investigator must set his standard before he obtains his results. Otherwise his choice may be influenced by his desire to obtain a certain verdict, of significance or nonsignificance.

2. The investigator should know what is implied by his standard of significance, and particularly by the standards that biological and medical workers commonly use—the 5% level (P = 0.05) and the 1% level (P = 0.01). He should therefore be aware of the 2 types of error to which all judgments are liable.

TWO TYPES OF ERROR IN JUDGMENT

In any judgment regarding the identity or difference of 2 things or phenomena, say A and B, one can make 2 kinds of mistake. One

can judge that A and B are different when they are really alike, or one can judge that they are alike when they are really different. Applying this to significance testing, we can define the 2 kinds of error thus:

A Type I error or "Error of the First Kind" is committed when there is no real (population) difference, but an observed difference (e.g., between the means of 2 samples) is proclaimed "significant."

A Type II error or "Error of the Second Kind" is committed when there is a real (population) difference, but an observed difference is proclaimed "not significant."

If, throughout a lifetime of investigation, a research worker sets his standard of significance at the 5% level, he insures that his Type I errors will not be greater than 5%; i.e., out of all the occasions on which chance alone is responsible for the differences that he observes, he will make an erroneous judgment (that something else is responsible) in not more than 5%.

It is impossible to set a general standard of significance that will insure a Type II error of a particular magnitude, but in the discussion of sizes of samples (p. 207) it will be seen that this error can be limited in particular cases.

By using the 1% level of significance, Type I errors of judgment are reduced to 1%; but this will in general increase the risk of Type II errors, for if we demand a larger difference before we proclaim it significant we shall be more likely to overlook real, but smaller, differences.

STATISTICAL TESTS AS A BASIS FOR ACTION

A strange sense of the unreal or theoretical seems to be attached in many investigators' minds to the choice of significance levels; but surely nothing could be more "real" or "practical" than the calculation of the risk of wrong or futile actions. When a laboratory worker asks such a question as "What is the proper value of P to rule out chance?" appropriate answers would be: "What risk are you willing to take of following a false clue? If, although you do not know it, something that you seem to have found does not really exist, how willing are you to go on hunting for it by further experiments? How willing are you to base a belief, and consequent action, on a nonexistent foundation?" To a clinician comparing a new treatment B with the customary treatment A, one could say: "What risk are you willing to take of changing to B when there is no real difference? If in all such therapeutic tests you adopt the 5% level of significance you will change from A to B 5% of the times when there is no difference between them. If you adopt the 10% level you will change 10% of the times, and so on."

The investigator's decision in each case must obviously depend on many factors, such as auxiliary information, the magnitude and importance of differences if they really exist, other advantages or disadvantages of clinical treatments (1) and, in general, the risk of Type II errors.

ANALOGY WITH ASSESSMENT OF CLINICAL NORMALITY

A physician should have no difficulty in seeing the analogy between the 2 types of error in significance testing and the errors that he risks in deciding whether a subject's weight, blood pressure, or other quantitative assessment is within clinically "normal" limits. For example, given the weights of healthy boys of a certain age, he may adopt as his standard the 10th percentile, i.e., the weight that cuts off the lightest 10% of the series. Applying this standard to a particular patient, a boy whom he is examining, if the boy's weight is below the 10th percentile he will classify him as "underweight." This is equivalent to a verdict of "significance" in a statistical test, and, as in that test, he may have committed a Type I error. That is, the boy's weight may not be low because of pathological factors, since 10% of the healthy boys had weights below the 10th percentile; but the physician rightly wishes to avoid overlooking a weight that is low from pathological causes. That is why he does not set the extreme lowest weight found in healthy boys as his lower limit for normality. If he did, he would pass a large number of patients whose weights were pathologically low; i.e., he would commit a Type II error very frequently.

SPECIFICATION OF SIGNIFICANCE LEVELS

A report should always indicate what level of significance is being adopted; and, more important, it should state the probability P for each result in such forms as "$t = 2.9$; P between 0.02 and 0.01"; "chi-square = 3.02; P between 0.5 and 0.3." The reader can then judge the evidence for himself and apply his own standard of significance.

Reports often contain such statements as "Although the difference was not statistically significant at the 5% level, it suggested a tendency for A to be greater than B," or ".... it seemed to indicate that A was greater than B." Such phrases commonly show that the investigator is adopting a lower standard of significance than the 5% level. The test of a belief is action, and in this instance if the investigator's actions are influenced in any way by the observed difference between A and B, e.g., in setting up another experiment to compare these 2 things, or in planning another part of the investigation, or in forming a theory, however tentative, re-

garding the phenomena that he is studying, then he must have some degree of belief that the observed difference was probably due to more than chance; i.e., his implicit standard of significance is less stringent than the 5% level.

REFERENCE

1. Mainland, D.: *Elementary Medical Statistics: The Principles of Quantitative Medicine* (Philadelphia: W. B. Saunders Company, 1952).

STANDARD DEVIATIONS AND STANDARD ERRORS

DONALD MAINLAND

EVEN RESEARCH workers who are familiar with elementary statistical techniques are, apparently, often confused by the terms "standard deviation" and "standard error." A statement like "Mean of sample of 16 measurements = 4.60 units; standard deviation = ±1.32; standard error of mean = ±0.33" is a kind of shorthand that many workers use with a sense of mystery. The lack of understanding may not impede the application of a significance test, but is serious when it leads investigators to ask: "Shall I test these averages by a standard deviation or a standard error?"

The first point to note is that "error" here is equivalent to "deviation," and both indicate variation or difference. It has become customary to use "standard deviation" with reference to the variation among individual measurements, and "standard error" with reference to variation among quantities such as means, that are derived from the individual measurements; but it is perfectly legitimate to use "standard deviation" throughout, and this usage is preferable because it removes some of the misunderstanding.

Some more of the mystery may be removed by reference to a disk sampling experiment (p. 127) in which each disk bears a measurement, such as a person's stature or blood pressure reading. Referring to the sample of 16 measurements, we may note 8 points:

1. The standard deviation in that example indicates the variation among the series of individual measurements, and is therefore best called the "standard deviation of individual measurements" or "standard deviation of the series."

2. The standard deviation of the individual measurements, calculated from the observed sample, is *an estimate* of the standard deviation that would be found in the population of individual measurements *randomly represented by the observed sample of measurements*. That is why, in calculating the standard deviation, if there are N measurements in the sample, $N - 1$ is used as divisor of the sum of squares of deviations from the mean; for it can be shown experimentally that this gives a better estimate of the true (population) standard deviation than does division by N itself.

3. If in a disk sampling experiment random samples containing, say, 16 disks per sample are taken, and the mean for each sample is found, until a thousand or more such means have been accumulated, there is created a population of means. The variation among

these means can be represented by a *standard deviation* calculated in the usual way, just as if the means were single measurements.

4. The "standard error of the mean" derived from an observed sample is better called a "standard deviation of means," and is an *estimate* of the quantity that would be found by the sampling in (3) above; i.e., it is an estimate of the standard deviation found in a population of *means of samples, of the same size as the observed sample, and randomly represented by the observed sample.*

5. In practice the estimate of the standard deviation of means is made by dividing the standard deviation of individual measurements (derived from the sample itself) by the square root of the sample size (e.g., $1.32/\sqrt{16} = 0.33$). This simple device can be justified experimentally; and even the fact that the sample gives only an estimate of the standard deviation of the original population of measurements is fully allowed for in such tests as the *t* test.

6. By another sampling experiment with a population of individual measurements one can take 2 samples (equal or unequal in size) each time, and find the difference between means of the 2 samples; or one can take 1 sample from 1 population and the other sample from another population. In either case one can accumulate a population of a thousand or more differences between means. The variation among these differences can again be expressed as a *standard deviation*, calculated as if the differences were individual measurements.

7. The "standard error of the difference" between 2 observed means (e.g., Difference = +2.30 units; standard error of difference = 0.89 unit) is better called a "standard deviation of differences," and is an *estimate* of the quantity that would be found by the sampling in (6) above; i.e., it is an estimate of the standard deviation that would be found in a population of *differences derived from samples, of the same size as the observed samples, and randomly represented by the observed samples.*

8. As in estimating the standard deviations of means, a simple device, justified experimentally, enables one to estimate the standard deviation of differences from the 2 observed samples themselves by the familiar method: Square the standard deviation of each mean, add the 2 squares and find the square root.

Note.—Whether an observer grasps fully the implications of such quantities as the standard deviations or not, it is his duty to make clear to his reader what quantities he is using. He should never leave the possibility of doubt as to whether a standard deviation (standard error) refers to individuals, means or other quantities. Such an expression as "4.60 ± 0.33" is indefensible unless it is shown what the quantity after the "±" sign really is. (This sign itself is, indeed, seldom necessary.)

SAMPLE SIZES

DONALD MAINLAND *and* LEE HERRERA, *New York University*

STATISTICIANS' ATTITUDE TO SMALL SAMPLES

SOME MISUNDERSTANDINGS still persist regarding the attitude of statisticians to small samples. For instance, Hill (2) discusses a physician's remark that the original detailed and exact description of fragilitas ossium, being based on only 2 cases, would mean nothing to a statistician. On the contrary, a good description is the 1st requisite for statistical or any other scientific work, and this 1st description was important evidence of the occurrence of a disease. But as soon as further questions arose about the disease, e.g., sex incidence or age incidence, relation to occupation or other diseases, or the effects of treatment, clinicians and statisticians alike would demand more cases.

Again, if a certain new treatment is applied to a patient with a disease that has been invariably fatal before and the patient (who undoubtedly has the disease) recovers promptly and completely, a statistician, like anyone else, would recognize that something very unusual has happened—either "due to" or merely "after" the treatment. But again the next questions, regarding actual causal relationship, frequency of cures, or necessary conditions of cure, would entail more cases.

The opposite type of criticism regarding sample size is also heard—that a statistician's conclusions from a small sample (20 or 30 cases) should have no weight against the experience of clinicians with hundreds of cases. Such critics do not heed the physician who said: "We make the same mistake a thousand times and call it 'clinical experience.' " Nor do they fully appreciate the fact that a valid statistical inference (1) makes clear the "population" or "universe" to which it applies, and (2) allows for sample size in estimating the chance or random variation between samples from this population or universe.

Such misunderstandings are becoming progressively less common, and there is a corresponding increase in the posing of the important question: How many patients (or animals) shall I require in this investigation? This question will be discussed later in this chapter.

LARGE SAMPLES

In the planning of medical surveys there is still much devotion to very large samples, and if it appears possible to include all members

of a population (e.g., all schoolchildren in a certain town to ascertain the incidence of dental caries, or all veterans with peripheral nerve injuries in a follow-up study), some investigators seem to think that this method is superior to any sampling process. The following 7 points should, therefore, be noted:

1. For administrative purposes (e.g., provision of dental care for the children actually surveyed, or medical care for the veterans), when all the subjects are examined they form an actual (finite) population; but when used for generalization regarding subjects of the same kind, or for comparison with other groups, they are a sample, and allowance must be made for variation between samples.

2. Mere enlargement of a sample does not automatically reduce bias.

3. However large a sample may be contemplated in the investigation, a pilot study on a small, preferably random, sample is often very desirable in order to ascertain the feasibility of the larger study, to explore methods, and to provide a basis for estimation of sample size required in the main study. The sample size in the pilot study depends on the particular investigation.

4. Because differences in location (geographical areas, different hospitals or factories in the same area) are recognized as important, plans are not infrequently made to use all the available material (subjects or records) in 1 location and then proceed to another if the results seem to warrant it, and if further financial aid can be obtained. Obviously, more knowledge about variation, and a broader basis for generalization, can be reached by taking random samples of, say, 100 subjects in each of 10 locations instead of 1,000 subjects in 1 location. A pilot study in 1 location, followed by extension, on the same or larger scale, to other locations is usually the most desirable method.

5. A sample of 2 or 3 dozen patients, thoroughly examined and properly recorded by skilled and careful physicians, is manifestly more reliable than a sample of several hundreds obtained from routine clinical records or examined for the investigation by physicians of varied skill and standards of carefulness.

6. Enlargement of sample size is affected by a law of diminishing returns with regard to precision. Thus, starting with the standard deviation of individual measurements in a sample of 100, we divide it by $\sqrt{100}$, i.e., 10, to find the standard deviation of the mean for a sample of that size, which can be represented by SD_{100}. For a sample of 200, the standard deviation of the individual measurements is divided by $\sqrt{200}$. Assuming the same standard deviation of individuals, this gives $SD_{100}/\sqrt{2} = 0.707 \times SD_{100}$. Increasing the

sample size by hundreds, we therefore find:

Sample Size	S.D. of Mean
100	SD_{100}
200	$0.707 \times SD_{100}$
300	$0.577 \times SD_{100}$
400	$0.500 \times SD_{100}$
500	$0.447 \times SD_{100}$
600	$0.408 \times SD_{100}$
700	$0.378 \times SD_{100}$
800	$0.354 \times SD_{100}$
900	$0.333 \times SD_{100}$
1,000	$0.316 \times SD_{100}$

The larger the sample, the less the effect of adding another 100 individuals.

The same phenomenon is observed with binomial confidence limits. If the observed percentage (e.g., of mortality) is 10, the limits (95% confidence band) for population percentages are as follows:

Sample Size	Confidence Limits (%)	Range between Limits (%)
50	3.3–21.8	18.5
100	4.9–17.6	12.7
200	6.2–15.0	8.8
300	6.9–14.0	7.1
400	7.3–13.4	6.1
500	7.5–13.0	5.5
1,000	8.2–12.0	3.8

In each instance the investigator should consider whether the gain by adding another 100 or 200 cases will compensate for the added cost in time, effort, and money.

7. The distinction between small and large samples is a matter of degree; but confusion has arisen because tests such as the t test were introduced primarily to treat samples that were so small as to give very unreliable results when treated by methods previously available. For convenience, a sample size of 15 or 20 was often taken as the boundary between "small" and "large"; and investigators who now meet such a rule and wish to compare, say, the mean of a sample of 7 with the mean of a sample of 25, are puzzled. This difficulty disappears when it is known that the t test is applicable to samples of any size. Indeed, it is so applied in analysis of variance. In ordinary medical research there is seldom any need to replace the t test by the "large sample" method, however large the sample.

EQUALITY OF SAMPLE SIZE

Amount of Information.—In the simpler statistical tests, samples that are being compared need not be of equal size in order to avoid

bias. For any given total number of subjects, however, more information can be obtained from equal samples than from unequal samples. An example of what happens in the comparison of two proportions, when the samples are equal and when they are unequal, is the following:

Suppose 2 samples with 40 items in each, and let 1 sample contain 5% of items with attribute A and the other sample contain 25% A's. It will be found that chi-square (with Yates's correction) is 4.80 and has a probability P of chance occurrence less than 0.05. The difference, 20%, between the 2 proportions is, therefore, significant at the 5% level. However, if the sample containing 5% A's consists of only 20 items, while the sample containing 25% A's consists of 60 items (making a total of 80 items as before), then chi-square = 2.60, with a probability of chance occurrence more than 0.10; and it follows that the 20% difference between the 2 proportions is not significant at the 5% level.

For measurement data, the same principle is demonstrated by the following example:

Consider 2 samples, each containing 5 measurements, with a difference of 5 units between their means, and suppose that the standard deviation of measurements, derived from the 2 samples, is $\sqrt{10}$ units. In order to test the significance of the difference we should proceed as follows:

SD of mean for each sample = $\sqrt{10}/\sqrt{5} = \sqrt{2}$ units (the same for each sample because the sizes are the same)

SD of difference between means = $\sqrt{2 + 2} = \sqrt{4} = 2$ units

Therefore

$$t = 5/2 = 2.5$$

Entering the table of t, with degrees of freedom = $(5 - 1) + (5 - 1) = 8$, we find that $t = 2.5$ has a probability P less than 0.05. The difference between the 2 means is, therefore, significant at the 5% level.

The situation is different, however, if 1 of the samples contains only 3 measurements while the other contains 7. Then, if the standard deviation of the measurements is the same as before, namely $\sqrt{10}$, the test proceeds as follows:

SD of mean of sample of 3 measurements = $\sqrt{10}/\sqrt{3} = \sqrt{3.33}$ units

SD of mean of sample of 7 measurements = $\sqrt{10}/\sqrt{7} = \sqrt{1.43}$ units

SD of difference between means = $\sqrt{3.33 + 1.43} = 2.18$ units

Therefore

$$t = 5/2.18 = 2.29$$

Degrees of freedom = $(3 - 1) + (7 - 1) = 8$ as before

The probability P is now greater than 0.05; and thus, although neither the difference between the means nor the variation between the individual measurements has been changed, we obtain a nonsignificant result.

Computational Advantage.—Arithmetically, also, there is an advantage in equal samples. Thus, when the means of a number of samples are compared by analysis of variance, the computation is simpler when the samples are equal in size; and when 2 percentages from equal samples are to be compared, there are tables available that entail little or no calculation by the user (4, 3). When more than 1 criterion of classification are used (e.g., sex and species, or 2 dose levels of 2 or more dietary supplements), the arithmetic for the treatment of unequal samples, i.e., unequal numbers in the various subclasses, becomes involved. Some methods for measurement data and also for frequency data are available (6), but an expert's help is often necessary.

When more than 2 criteria of classification are used there are not many cases in which a satisfactory method is available for inclusion of all the data in 1 analysis without risk of bias; and expert advice is necessary. Great difficulties, therefore, often arise in data obtained from surveys, in contrast to experimental data. An experiment can, of course, produce the same kind of difficulty by irreplaceable loss of subjects; but the experimenter can at least avoid purposely introducing inequality by adding 2 or 3 extra subjects in certain subclasses, or extra observations on a few of the subjects. The most economical way of treating such data is often to throw out, by strictly random selection, all the extra observations.

SIZES OF SAMPLES REQUIRED

The question "How large should my sample be?" generally means 1 or both of 2 things, expressed here in terms of the comparison of 2 samples:

1. If a real difference exists, how large must my samples be in order that the difference found in my experiment will be statistically significant?

2. If there is no real difference, how large must my samples be in order to prove that there is no real difference?

Question (2), in this form, cannot be answered, for it is impossible to prove that no difference exists. By increasing sample size one can merely narrow the confidence limits, on each side of zero, between which the true (population) difference probably lies.

As a basis for answering the 1st question, one requires some information, from one's own previous observations or those of others, regarding the magnitudes of differences and the variation likely to be met among subjects treated alike; or, as a much less satisfactory

alternative, one must make a number of suppositions and estimate sample sizes based on each supposition.

A simple and commonly used method is to take the values found in a preliminary investigation, e.g., a difference between means and the standard deviation of the difference, and to assume that the difference, and also the variation between individuals, would be the same in the larger investigation. Use is then made of the well-known relationship between the sample size and the standard deviation of the mean; e.g., quadrupling the sample size will halve the standard deviation. By a cognate process with frequency data, such as deaths or recoveries, if the proportions in a fourfold table remain the same, doubling the sample size will double the value of chi-square. (When chi-square is computed with Yates's correction, this relationship is not quite exact, but sufficiently close.)

When sample sizes required to meet a specified level of significance are estimated by these methods, they are useful if they are found to be so large as to show that the investigation is impracticable; but the underlying assumptions are too limited for other purposes. Thus, the difference between means, or between frequencies, in the enlarged experiment might well be smaller, or the individual variation larger, than in the original experiment; and then the difference would not reach the required level of significance. In the long run, therefore, such estimates provide only about a 50:50 chance of success. To be somewhat safer, the observer usually takes samples rather larger than these minimal estimates; but it is desirable to know how large they should be in order to justify much greater confidence of success than 50%.

It is now possible to make such estimates with precision if the investigator, inquiring about sample size, answers 3 questions which, for ease of discussion, can be phrased in terms of 2 treatments, V and W, each applied to an equal sample of subjects, the effects being recorded either as measurements or as frequencies. The questions are:

1. If there is actually no real (population) difference between the effects of V and W, what risk are you willing to run of mistakenly concluding that there is a difference?

2. If there is a real difference between the effects of V and W, what size of difference is important to you?

3. If there is a real difference of the size specified in answer to Question (2), what risk are you willing to run of failing to detect a difference?

Reference to the 2 kinds of error in judgment (p. 196) should, along with the following remarks, elucidate these questions.

Erroneous inference of a difference.—This is a Type I error, and it

s customary in biology and medicine to set the maximum allowable error of this type at 5%. For greater assurance the investigator may adopt the 1% level of significance, and then, of course, the estimated sample size will be greater.

Size of a real (population) difference.—If the investigator said that any real difference, however small, would be important to him, he would commit himself to ever larger and larger samples. What is necessary is a statement such as the following: "If treatment V would in reality lower the blood sugar on the average by 20 mg per 100 ml more than treatment W, I wish to have considerable confidence that my experiment will be successful, i.e., that it will reveal a greater effectiveness of the V treatment by producing a statistically significant difference." A similar statement could be phrased in terms of frequencies. Thus, the stipulation might be that if treatment V would reduce a population percentage mortality by 10 (e.g., 40% with W, 30% with V), the difference found in the experiment should be significant.

NOTE.—These statements do not imply that the difference observed in the experiment should be 20 mg in the 1 case or 10% in the other; but merely that the observed differences, whether greater or less than these population values, will be great enough to pass the test of significance.

Failure to detect a real difference.—This is a Type II error, and the risk of this kind of error can be set at 1%, 5%, or higher, according to the wishes and needs of the experimenter. Such statements in terms of error can be expressed also as degrees of confidence in a "successful experiment," i.e., 1 that will provide a significant difference, or as a statement of probability. Thus, if the risk of error is set at 10%, the investigator will have confidence that in a long series of experiments, under the specified conditions, 90% would be successful; i.e., the probability of success in any 1 experiment is 90%.

Sources of information.—When the investigator has answered the foregoing 3 questions he can obtain an answer to his question on sample size. For measurement data he may need some personal assistance in using the techniques (1). For frequency data (comparison of 2 percentages) he can read the answer directly from tables for sample sizes up to 100 (5). (Tables for larger samples are not yet prepared.)

NOTE.—In planning a long experiment where loss of subjects might seriously jeopardize the outcome, it is very advisable to make allowance for such loss, even to the extent of finding an upper confidence limit for percentage mortality estimated from appropriate previous data.

REFERENCES

1. Harris, M.; Horvitz, D. G., and Mood, A. M.: On the determination of sample sizes in designing experiments, J. Am. Stat. A. 43: 391, 1948.
2. Hill, A. B.: The clinical trial, New England J. Med. 247: 113, 1952.
3. Mainland, D.; Herrera, L., and Sutcliffe, M. I.: *Statistical Tables for Use with Binomial Samples in Medicine and Biology*, in preparation.
4. Mainland, D., and Murray, I. M.: Tables for use in fourfold contingency tests, Science 116: 591, 1952.
5. Mainland, D., and Sutcliffe, M. I.: Statistical methods in medical research: II. Sample sizes required in experiments involving all-or-none response. Canad. J. M. Sc. 31: 406, 1953.
6. Snedecor, G. W.: *Statistical Methods Applied to Experiments in Agriculture and Biology* (Ames, Ia.: Iowa State College Press, 1946).

NONMETRICAL TESTS OF MEASUREMENT DATA

DONALD MAINLAND

When measurements have been made in an experiment, it is customary to use the actual measurements in testing the results, e.g., in testing a series of differences by *t*. Not infrequently, however, the experimenter could obtain the information that he desires by a quicker and simpler method, and it is noteworthy that Fisher (1) gave an example of this along with his 1st illustration of the *t* test. Of the 10 patients treated by 2 supposed soporifics (p. 192), 1 showed no difference between the dextro- and the levorotatory isomers with regard to the hours of sleep gained; the other 9 all gained more when they received the levo- compound. That is, when the difference was expressed as levo- minus dextro-, there were 9 plus signs, and the test now to be described is called the *sign test*.

THE SIGN TEST

The patient who showed no difference gave no evidence in either direction and was therefore excluded from the test. (In reality a zero difference indicates that the difference, whether plus or minus, is too small to be detected by the method of measurement that is used.) The question then is: If in the long run the number of patients with plus signs (superiority of the levo- compound) were equal to the number with minus signs (superiority of the dextro-compound), how often, by chance, would one meet samples that departed as far from this 50:50 ratio as did the sample of 9 patients? The answer is obtained from the expansion of the binomial $(0.5 + 0.5)^9$ and shows that such extreme samples (with all plus signs or all minus signs) would form 0.4% of the total samples, i.e., P, the probability of chance occurrence, is less than 0.01, and the departure from the 50:50 ratio is significant at the 1% level. This is the same verdict as was given by the *t* test.

Answers to the sign test can be obtained directly from tables of binomial confidence limits (2, 3). For example, with 1 minus and 8 plus values, the number of *A*'s in the sample is 1, and when $N = 9$ the upper limit (95% band) is 48.3% *A*'s, i.e., less than 50%. There is a "significant minority" of minus values (P less than 0.05). The upper limit in the 99% band, however, is 58.6%, i.e., greater than 50%; and the difference from 50% is therefore not significant at the 1% level.

Returning to the actual experiment, with no minus values in a sample of 9, we use the table for No. of A's in sample = 0, and find that the required upper limit is 44.5% (in the column headed "P = 0.005," which corresponds to the 99% limit in the tables where A is greater than zero). The verdict, as obtained from the binomial expansion, is that there is a significant minority of minus values (P less than 0.01).

The sign test and the t test do not invariably give the same verdict, for they are not asking exactly the same question. The t test is concerned with the mean of differences that are, at least approximately, normally distributed. The sign test is concerned merely with the direction, plus or minus, of the differences. It is therefore less sensitive; i.e., it may not show a significant minority of either sign, whereas, if the individual values that contribute to the mean agree closely with each other, the t test may show that the mean is significantly different from zero on the plus (or minus) side. The sign test, then, is less efficient than the t test, in that it sacrifices information about actual magnitudes.

Occasionally, however, the verdicts are reversed. For example, if there is 1 negative difference and 9 positives, the upper limit (95% band) is 44.5%; there is a significant minority of negative values. But if the negative value is large and the positive values are relatively very small, the t test may show a mean not significantly different from zero. If the observations are reliable, such an occurrence suggests that the distribution of the values, if more were obtained, would be very skew, or that there are 2 classes of subjects, a majority who react positively but not strongly, and a minority who show a strong negative reaction. Further exploration is obviously desirable.

OTHER NONPARAMETRIC TESTS

Such tests as the sign test are called "nonparametric" because they do not entail the estimation of such quantities as means and standard deviations, which are "parameters," i.e., measures of certain features of the population to which the data belong.

A nonparametric substitute for the comparison of means is, like the sign test, less sensitive than the t test with approximately normal distributions, but it is very useful with some kinds of data. When the effect of a treatment is measured by the duration of symptoms or signs (e.g., number of days before pus ceases to form, or before temperature reaches normal levels), 1 or 2 patients in a series often take much longer than the rest to reach the specified state; that is, the distribution is very skew. If in the comparison of 2 treatments these outlying cases are omitted, bias results, whereas

if they are included in, say, a comparison of mean duration of symptoms, the difference between means may be found not significant when in reality 1 treatment is superior to another. The simplest form of nonparametric analysis can be illustrated by supposing that 20 patients have been treated by method *V* and 20 by method *W* (but sample sizes need not be equal). Arrange the patients in ascending order of the variate, say days to recovery, and find the median (or a close approximation to it), i.e., the value, say 10 days, that divides the total 40 patients into 2 halves. Arrange the data in a fourfold table thus:

	Days to Recovery	
	Less than 10	10 or more
Treatment *V*	...	...
Treatment *W*	...	...

The difference in proportions can then be tested by chi-square or directly from tables (4, 3).

When samples are large enough, the same technique can be extended by dividing the data not only at the median (the 50% point) but at intermediate points, such as 25 and 75% or at 10, 20, 30%, and so on. This increases the number of columns in the table, and, in order that the chi-square test can be applied without risk of error from small numbers, the total of each column (when the *V* and *W* samples are equal) should usually be not less than 10 subjects.

The same methods are useful with incompletely metrical data such as immunization titers, which often contain some indeterminate values, e.g., "less than 0.002" and "greater than 0.5."

A nonparametric correlation test is the method of "rank correlation," of which the elementary technique is well shown by Rosander (6); but for guidance in the choice and application of the method, experimenters should seek personal help. Investigators who already have some familiarity with statistical techniques will find a useful survey of nonparametric methods in an article by Moses (5).

REFERENCES

1. Fisher, R. A.: *Statistical Methods for Research Workers* (Edinburgh and London: Oliver & Boyd, Ltd., 1925).
2. Mainland, D.: *Elementary Medical Statistics: The Principles of Quantitative Medicine* (Philadelphia: W. B. Saunders Company, 1952).
3. Mainland, D.; Herrera, L., and Sutcliffe, M. I.: *Statistical Tables for Use with Binomial Samples in Medicine and Biology*, in preparation.
4. Mainland, D., and Murray, I. M.: Tables for use in fourfold contingency tests, Science 116: 591, 1952.
5. Moses, L. E.: Non-parametric statistics for psychological research, Psychol. Bull. 49: 122, 1952.
6. Rosander, A. C.: *Elementary Principles of Statistics* (New York: D. Van Nostrand Company, Inc., 1951).

CONSULTATION WITH A STATISTICIAN

DONALD MAINLAND

THE TERM "statistics" covers such a wide variety of specialized activities that a medical research worker, unaware of this, may go far astray in choosing a statistician to help him. He should, of course, not expect help from a mathematician, even from one who specializes in statistical mathematics, unless he can find one who has had experience in the application of statistics to the kind of problem that is involved. He may sometimes find an experimental statistician in a department of preventive medicine or public health, but many statisticians in such departments can give assistance only in their own field. Often the best source of help in planning and analyzing an experiment is a worker in applied biology (e.g., experimental agriculture), but he is not the one to give expert assistance in avoiding the pitfalls of vital statistics.

Some hints may help to reduce waste of time and to minimize misunderstanding between the investigator and the statistician whom he consults. For brevity the suggestions will be enumerated as 10 rules:

1. Because of the rapid expansion of medical statistics and the great scarcity of statisticians, for every project to which they can give attention there are many projects for which they can find no time at all. Be patient when your problem does not receive attention as quickly as you would desire, or if assistance has to be declined.

2. Interruption in a statistical analysis may necessitate the repetition of several hours' work. Try to make an appointment in advance.

3. If you have been immersed in a problem for weeks or months do not expect a statistician to grasp it in half an hour well enough to give a safe answer to your questions. It saves much time if, before your visit, you send a page or two containing the following information:

a) A specific statement of the object of your investigation—what question or questions it is intended to answer.

b) An outline of the method, including numbers of subjects.

c) A sample of results, if any have been obtained.

d) A precise statement of what you wish the statistician to do.

4. As is stressed throughout this Section, the proper time to

consult a statistician is in the planning stage. If you have already done the investigation and merely want a statistical analysis, unless you have planned and conducted the work properly do not expect the statistician to do anything with the data except indicate how a proper investigation should be conducted. A statistician who analyses data with which he is not satisfied is doing a disservice to science, to the investigator, and to his own reputation.

5. Even when a proposed experiment or survey appears very simple, do not be surprised if the statistician points out some great, or even insuperable, difficulties.

6. If the statistician points out a flaw in the research method or an invalid inference, accept his statement as unlikely to be wrong, even if you do not understand it.

7. Seek thoroughly for possible weaknesses in your work which the statistician, through ignorance of the investigational technique, may not be aware of; and discuss them with him.

8. If you have already collected some data it is usually best to present them to the statistician in their original "raw" form. Do not invent complicated expressions (ratios and the like) and expect them to be suitable for statistical tests. Even workers who know some orthodox techniques, such as standard errors, often waste time by applying them where they are inappropriate.

9. If the investigation is a good one, it is worth the time and money spent on the necessary computation. Budget for it as for the rest of the investigation.

10. Because any investigation may contain features for which the statistician does not wish, by implication, to assume responsibility, do not give him any acknowledgment, either orally or in print, unless he has previously approved of its actual wording.

Comment by E. Cuyler Hammond

This Section should be required reading for all research workers in the biological sciences, not excluding statisticians. In making this flat statement I must plead guilty to a certain amount of prejudice, since one is prone to like any publication which emphasizes the importance of his own profession.

A scientific investigation is carried out in 3 successive stages: (1) design of the experiment; (2) collection of data, and (3) analysis of results. This Section is concerned almost entirely with the first 2 stages, which are by all odds the most important, which are the primary concern of laboratory and clinical investigators, and which are largely neglected in most textbooks on statistics. As to the technical statistical analysis of his data, the reader is left the choice of consulting a textbook or consulting a statistician. However, the statistician can be of most assistance if he is first consulted in the planning stages rather than later.

Comment by S. Lee Crump

This Section should help to dispel many misconceptions which are prevalent among researchers in the medical fields about the role and nature of statistics as currently viewed. Anyone associated with medical research will find time devoted to reading it well spent. It is not a textbook of statistical methodology.

SECTION IV

Design and Construction of Metabolism Cages

ASSOCIATE EDITOR—*Arnold Lazarow*

INTRODUCTION

In order to carry out metabolic studies in animals it has been necessary to design cages which would permit the quantitative measurement of both the food and water intake as well as the urine and feces output. It is obvious that the cage design will vary with the type of animal used, and with the specific requirements of the experiment. In many dietary experiments, for example, the quantitation of the food intake may be of paramount importance; hence special precautions are required to minimize spillage. In studies on experimental diabetes, the quantitative collection of urine may be the prime concern, whereas in studies with radioactive isotopes, the quantitative collection of the exhaled CO_2 may be necessary.

In the sections which follow, the cage design for the various species of animals (rat, mouse, dog, and monkey) will be discussed individually. The special problems concerned in measuring the intake of food or water and the output of urine, feces, or CO_2 will be considered separately. Special consideration will be given to the problem of cage maintenance, efficiency of cleaning and avoidance of contamination.

—A. Lazarow.

A. RAT METABOLISM CAGES

ARNOLD LAZAROW, *Western Reserve University*

METABOLISM CAGES designed for the quantitative collection of urine and feces have been in use for a half-century. In 1904, Henriques and Hansen(9) described a metabolism unit in which a round wire cage was set in a large glass funnel. The urine was separated from the feces by a screen placed within the funnel. Although a number of modifications have been described (1, 5, 7, 8, 18, 21, 24), most of the modifications embody the original cage design of Henriques and Hansen. We have recently devised 2 types of square metabolism cage units (15) which are suspended from a rack and hence are extremely compact and much more convenient to use.

Animals kept on screen-bottomed cages are best maintained at a temperature of 68–72 F and a relative humidity of 50%. Air conditioning is important inasmuch as bedding cannot be used in metabolism cages.

I. Round Wire-Mesh-Glass Funnel Rat Metabolism Cage*

DETAILS OF CONSTRUCTION

The metabolism unit shown in Figure 1 has been in common use in many laboratories. Essentially it consists of a round cage *A*, made of $^1/_2$ in. wire mesh. The cage is set into a glass funnel *B* which in turn is mounted in the wooden frame *C*. The legs *D* rest on the surface of the rack and help support the cage. The stem of the funnel *E* passes through the hole *F* and into the urine collection bottle *G*. The wooden block *H* is used to bring the mouth of the urine collection bottle around the stem of the funnel. A circular screen *J* made of $^1/_4$ in. wire mesh is placed within the funnel *B*. The feces pass through the wire mesh screen which forms the bottom of the cage, and are retained by the screen *J*. The urine passes through the screen *J*, through the filter paper *K*† and into the collection bottle *G*.

---

* Available through most laboratory supply houses or the George H. Wahman Company, Baltimore 2.

† Since most large funnels do not have an exact 60° angle, the filter paper should be folded so that point *L* is slightly below and to the right of point *M*. When this cone is opened the angle will be slightly more than 60°, and hence it can be made to fit the funnel more exactly. When the filter paper *K* is placed within the funnel *B* it is moistened with water to keep it in place.

An assembled unit with external drinking fountain and food cup is shown (*P*). The details of the drinking fountain *O* are shown in Figure 8, *A*, and of the food cup *N* in Figure 11. An assembled unit with an internal food cup and drinking fountain is shown in *S* (Fig. 1). The non-scatter food cup *Q* is shown in Figure 10, *A*, and the drinking fountain *R* in Figure 9.

The single metabolism unit shown in Figure 2 consists of a body *A* made of $^1/_2$ in. wire mesh; the bottom of the cage is fabricated into a metal funnel *B*. The 3 legs *C* which attach to the base *D*

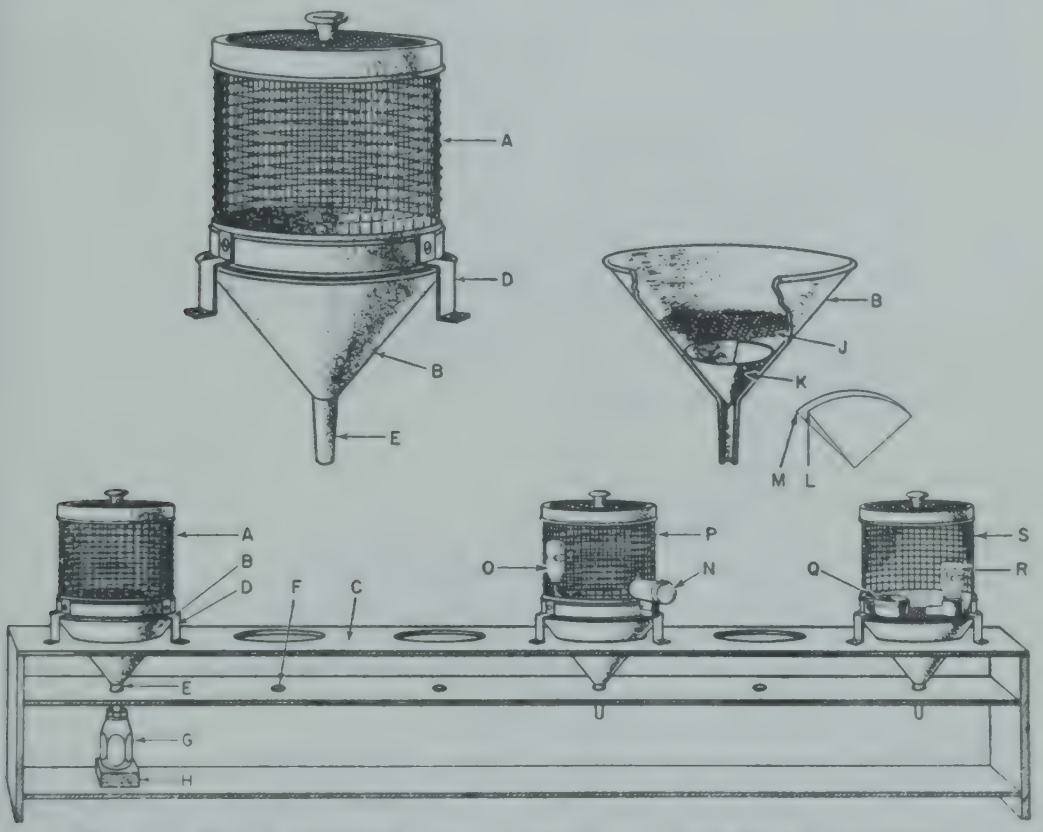


FIG. 1.—Rat metabolism cage: round cage-glass funnel assembly.

serve to support the cage. A false bottom (not illustrated) made of circular wire screen fits into the funnel *B*.

Comment.—The extent to which these cages have been used bears testimony to their utility. However, the cage-funnel assembly shown in Figure 1 is bulky and requires considerable table or shelf space. It is, furthermore, difficult to transport the assembled metabolic unit.

Many models, which are available commercially, do not have the legs (*D*, Fig. 1) and hence are less stable. Accidental jarring, vigorous movement, or a convulsion may be sufficient to upset the cage. The over-all stability is increased by having the legs of the

cage rest on the top surface of the rack. However, this addition necessitates that the dimensions of the rack be exactly tailored to the dimensions of the funnel; the projection of the funnel above the top shelf of the rack must be less than the height of the legs. The hole *F* in the middle shelf of the rack should be only slightly larger than the stem of the funnel *E*. Since the wooden support racks are usually not available commercially, they can be designed to fit the funnels that are available.

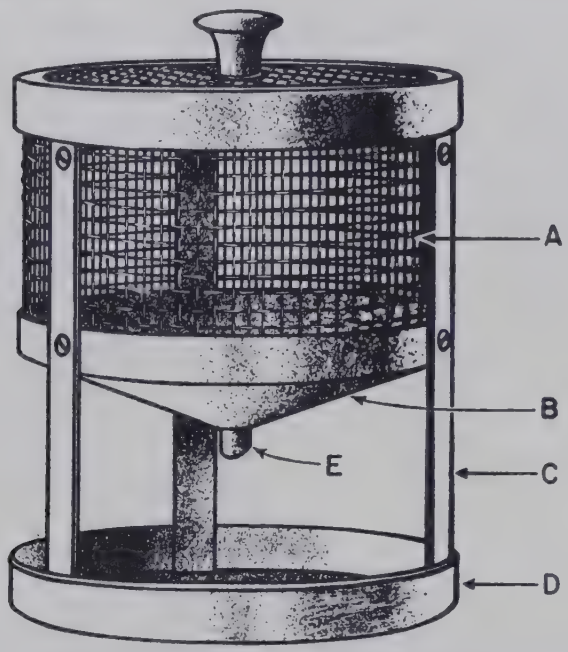


FIG. 2.—Rat metabolism cage: individual unit.

The glass funnels are expensive and easily broken. Tin-coated funnels may be used but they are often difficult to obtain in the appropriate size. Furthermore, tin-coated funnels tend to rust and corrode on continued use. We have been unable to locate a commercial source of aluminum or stainless steel funnels of appropriate dimensions.

In cleaning the metabolism unit the cage must be completely disassembled to clean the funnel and replace the filter paper.

The metabolism unit shown in Figure 2 may be convenient for single rat experiments, but such cages are cumbersome to use when a large number of animals are being studied. In order to clean the cage the rat must be removed, and fabrication of the funnels into the cage certainly does not facilitate the cleaning procedure.

II. Suspended Rat Metabolism Cage‡

DETAILS OF CONSTRUCTION

This all stainless steel unit (Fig. 3) consists of a box 7 in. wide, 8 in. deep and 6 in. high. The sides *A*, top *B* and back are made of stainless steel sheet (no. 26 gauge). The lateral projections *C*, which are spot-welded to the sides of the cage, are made of 20

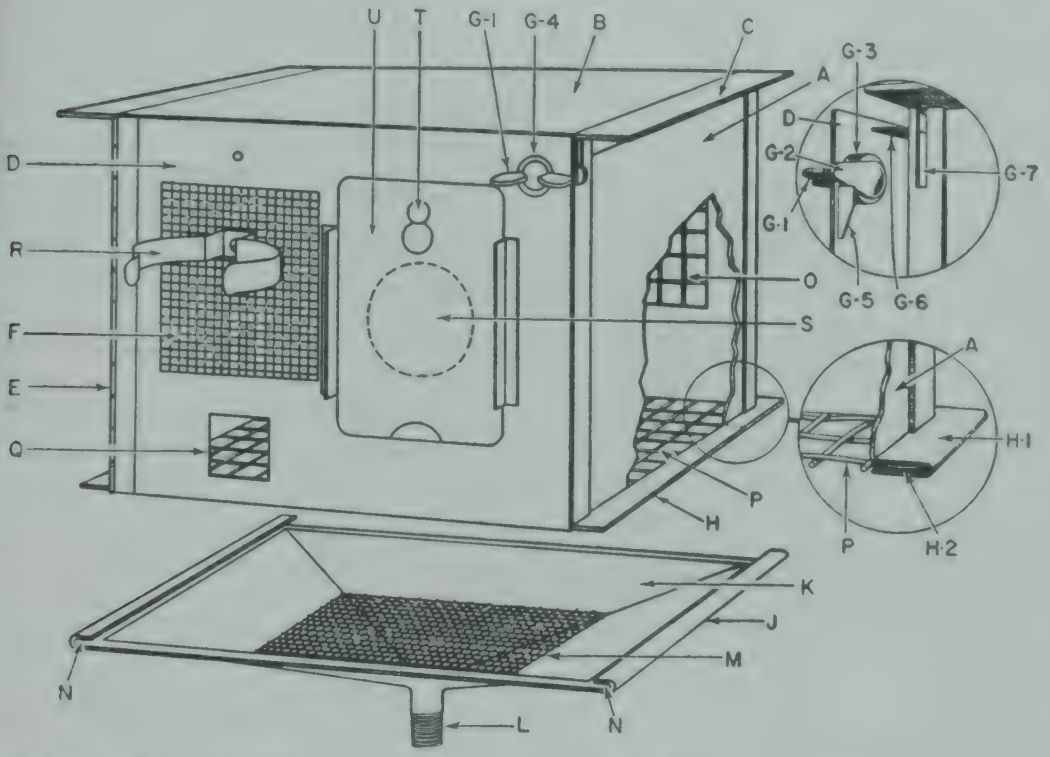


FIG. 3.—Rat metabolism cage: suspended type. Details on component parts

gauge steel and serve to support the cage in the rack. The bottom of the cage *P* consists of a stainless steel screen ($^{7}/_{16}$ in. mesh) soldered into the groove recess *H–2* which is formed by bending the walls of the cage as shown in the lower inset (*H–1*).

The door of the cage *D* is made of no. 18 gauge stainless steel plate. The hinge *E* is 6 in. long and attaches to the left side of the door. A portion of the door is replaced by a screen window *F*, which is made up of stainless steel wire ($^{1}/_{4}$ in. mesh). A spring clip *R* supports the drinking fountain (see Fig. 4). The upper two thirds of the back of the cage is likewise replaced by a screen window *O*, which is made of woven wire ($^{7}/_{16}$ in. mesh).

The rectangular opening *Q* is $^{7}/_{8}$ in. square and it accommodates

‡ Available from the Micro-Metric Instrument Co., P.O. Box 884, Cleveland 22.

the mouth of the drinking fountain. The round opening *S* is 2 in. in diameter and it permits the attachment of an external food cup. When a non-scatter food cup is used within the cage, a metal face plate *U* covers this opening. This face plate is secured to the door by the rivet *T*.

Details of the door lock are shown in the upper inset (Fig. 3). A wing bolt *G–1* is soldered into the metal tab *G–2*. These are separated from the door *D* by the washers *G–3* and *G–4*. The

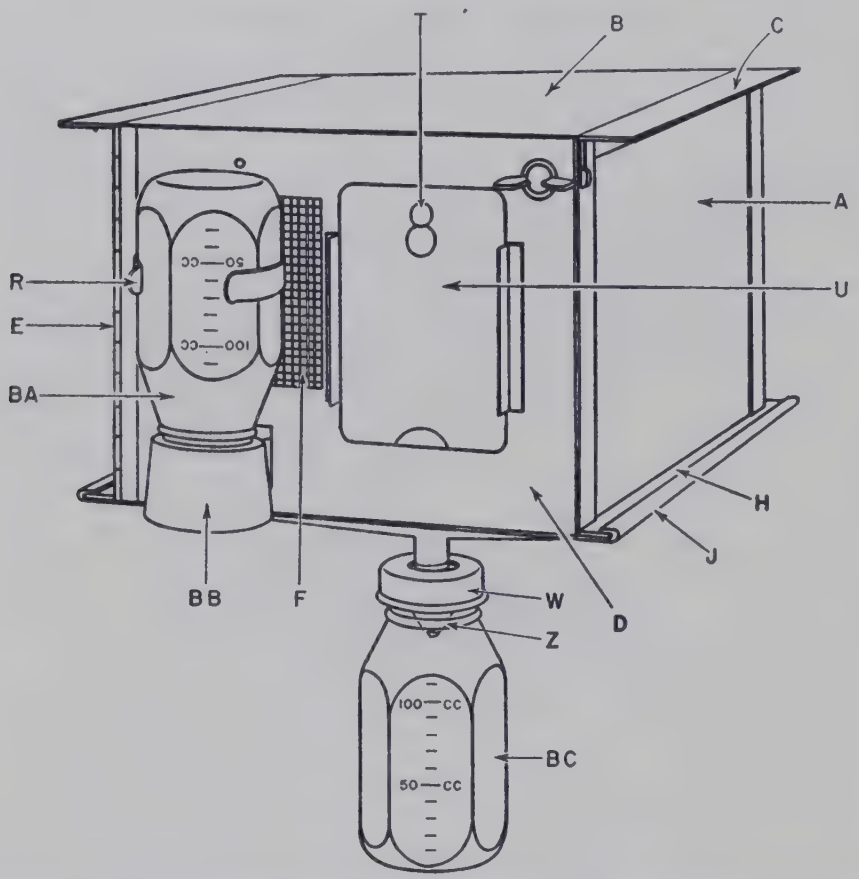


FIG. 4.—Rat metabolism cage: suspended type. Drinking fountain and urine collection system assembled.

metal stops *G–5* and *G–6* restrict the movement of the lock. The door is locked by turning the wing bolt a quarter turn. This brings the tab *G-2* into the groove *G–7*, which is milled into the side of the cage.

The funnel *K* is rectangular and has 4 sloping sides. The lateral margins of the funnel are bent to form the grooved channel *N*. Height of the funnel from base to apex is 2 in. A threaded nipple *L* ($^3/_8$ in. pipe thread) is soldered into the apex of the funnel. The funnel is attached to the cage by sliding the lateral margin of the funnel *J* over the lateral projections *H* which form the base of

the cage. A metal stop which is spot-welded to the back of the cage (not illustrated) prevents the funnel from sliding beyond the limits of the cage.

The assembled metabolism unit is shown in Figure 4. The animal drinking fountain *BB* employed is of the type illustrated in Figure 9. The bottle is supported by the spring clip *R*. A non-scatter food cup is used within the cage and the face plate *U* is in place. The urine collection system is of the type shown in Figure 13. The bottle *BC* with its contained funnel *Z* is attached to the large funnel by the cap *W*.

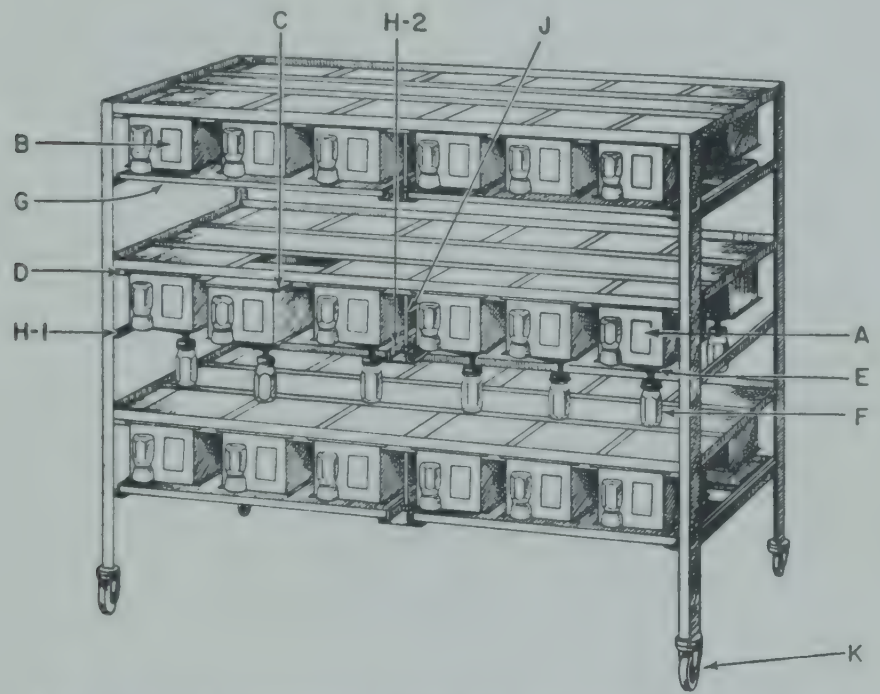


FIG. 5.—Rack containing suspended type of metabolism cage.

The rack and assembled metabolism units are shown in Figure 5. The rack measures $5\frac{1}{2}$ ft wide, 2 ft deep and 5 ft high; it is mounted on 4 in. casters *K*. The rack accommodates 3 double rows of cages, making a total of 36 units, 18 cages on each side. The cages are suspended in the rack by sliding the projections *C* into the appropriate groove *D* on the rack. The cages *A* in the second row of the rack are assembled with funnels *E* and urine collection bottles *F*. When quantitative urine collection is not required, the funnels and bottles are removed and the cages are assembled as shown in the top row (*B*). The feces and urine from 3 cages are collected in a single stainless steel pan *G*, 30 in. long, 10 in. wide and $\frac{3}{4}$ in. tall, supported by the angle irons *H-1* and *H-2* and the rods *J*.

Comment.—This all stainless steel unit has been in use in our

laboratory for the past 3 years. The convenience of construction has permitted us to carry out simultaneous metabolic studies on 144 rats. The cages are extremely compact and easy to clean. The rack containing 36 units can be moved about without danger of

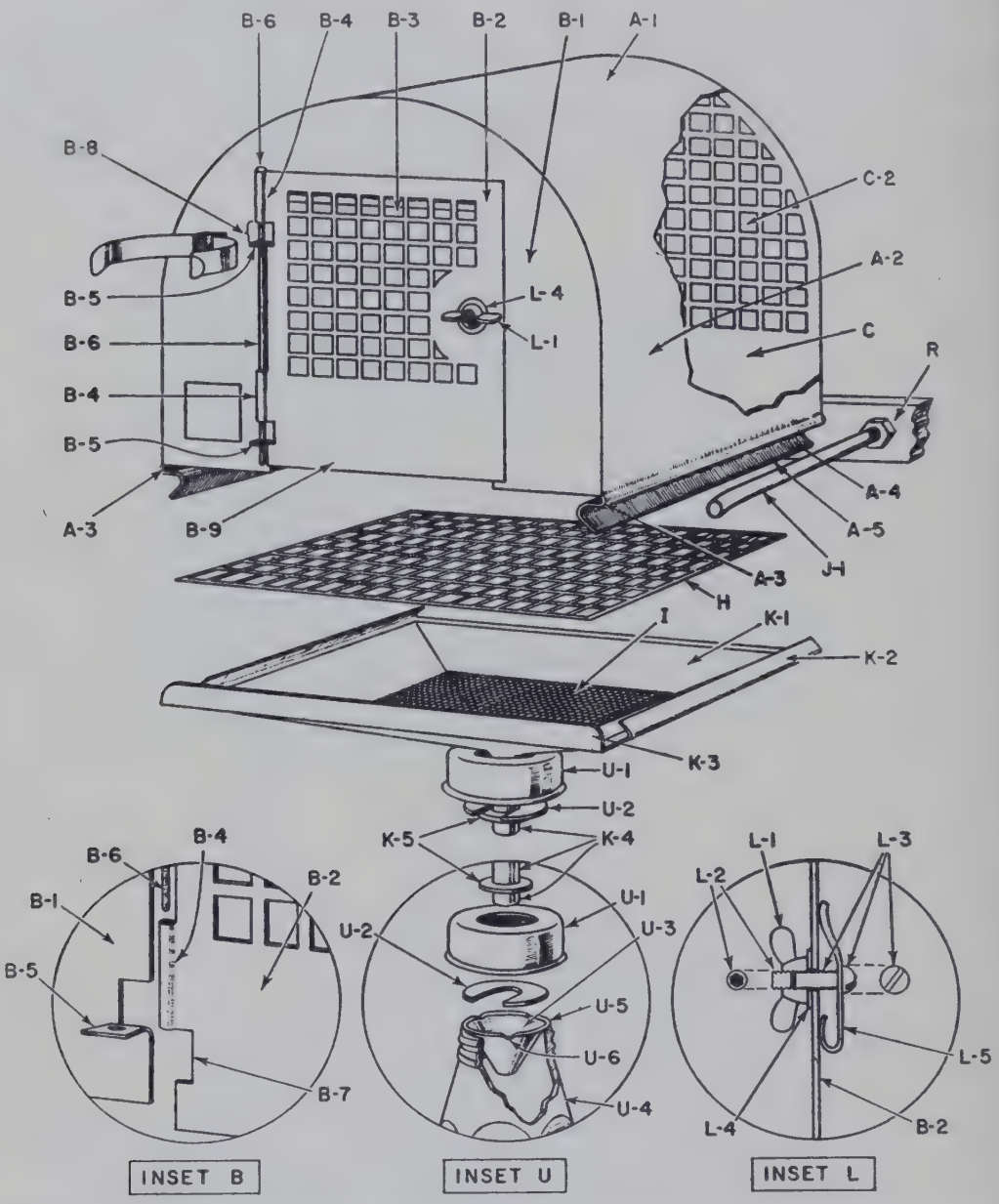


FIG. 6.—All-purpose rat metabolism cage; detailed view of component parts

spilling. The funnels can be removed for cleaning without disturbing the cage or the rat. The urine is filtered and collected in a calibrated bottle. During the days when metabolic measurements are not required, the individual funnels are removed and large pans are used to collect the urine and feces. When these cages are

used with radioactive materials, complete decontamination may be difficult because material may accumulate in crevices such as *H–2*, Figure 3.

III. Improved All-Purpose Rat Metabolism Cage Suitable for Radioisotope Work

The need for an all-purpose metabolism cage which could be easily assembled, easily cleaned, effectively decontaminated and inexpensively fabricated has led to the construction of the unit shown in Figures 6 and 7.

DETAILS OF CONSTRUCTION

This cage, made of stainless steel, has a rectangular base and cylindrical top. A single bent strip of metal forms the top *A–1* and the sides *A–2* of the cage; it is attached to the front *B–1* and back *C* of the cage by means of a "Pittsburgh seam." The lower margin of the side of the cage *A–4* is adapted to retain the false bottom *H*, to serve as a point of attachment for the urine collection funnel *K–1*, and to attach the cage to the rack *R*.

The upper two thirds of the back of the cage *C* is perforated to permit visibility. The holes *C–2* are $^{5}/_{16}$ in. square and have curved edges. The door of the cage *B–2* is likewise perforated by a series of square holes *B–3*.

The hinge is fabricated from the door and the front of the cage. Two rectangular tabs at the left margin of the door are bent into the hinge retainers *B–4* (see also inset *B*, Fig. 6). Two rectangular pieces of metal are removed at point *B–7* to provide clearance. The hinge supports *B–5* are formed by bending the 2 flaps of metal (inset *B*). The door is supported by inserting the stainless steel pin *B–6* through the tubular hinge retainers *B–4* and the hinge supports *B–5*. The door can be detached from the cage by withdrawing the pin *B–6*.

The lock (see inset *L*) consists of a bent metal piece *L–5* which is spot-welded to the screw *L–3*. The screw *L–3* is inserted through a hole in the door *B–2* and the washer *L–4*. The wing nut *L–1* is tightened until appropriate tension is obtained on the spring clip *L–5*. An Allen screw *L–2* is inserted into the wing nut *L–1* and tightened against the screw *L–3;* this firmly secures the wing nut on the screw *L–3*. The lock is closed by turning the wing nut *L–1* a quarter of a turn.

The false bottom of the cage is a stamping which is available in

§ Designed and constructed by the Micro-Metric Instrument Company, P.O. Box 884, Cleveland 22.

several sizes. For rats of 150–300 g, a $^{5}/_{16}$ in. square mesh is suitable. In order to insert this false bottom into the grooves *A–3*, it is necessary that the bottom of the door *B–9* be brought above the level of the groove *A–3*. This is accomplished by lifting the door *B–2* in a vertical direction until a clearance *B–8* develops between the hinge retainer *B–4* and the hinge support *B–5* (see Fig. 6). The false bottom *H* is inserted into the groove *A–3* and the door is

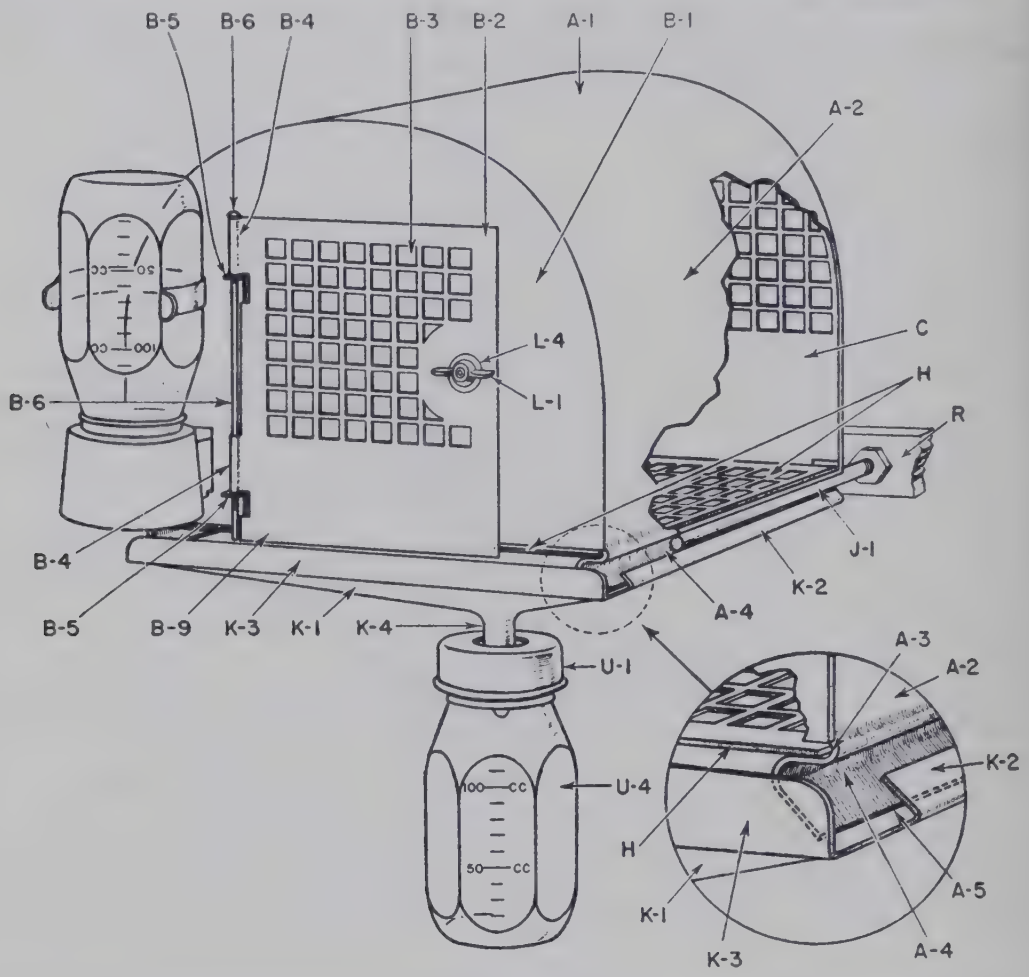


FIG. 7.—All-purpose rat metabolism cage; drinking fountain and urine collection system assembled.

lowered. When the space *B–8* between the hinge retainer *B–4* and the hinge support *B–5* is obliterated as shown in Figure 7, the false bottom is held in place within the cage by the lower portions of the pin *B–6* and door *B–9* which lie below the level of the groove *A–3*.

The urine collection funnel is a stainless steel spinning. The cone-shaped structure thus obtained is converted into a rectangular pyramid *K–1* by deforming it in a mold. The lateral borders of the funnel are cut and bent to form the grooved channel *K–2*. The front of the funnel is bent through 90° to form a flap *K–3*. The

stem of the funnel *K–4* is formed in the original spinning. A metal ring *K–5* (see inset *U*, Fig. 6) is brazed on the stem and serves to support the urine collection bottle. The funnel is attached to the cage by sliding the channel *K–2* over the lateral margin *A–5*, as shown in Figure 7. The front projection of the funnel *K–3* acts as a stop.

The urine collection system is shown in inset *U* (Fig. 6). The stem of the funnel *K–4* with its attached ring *K–5* are inserted through the plastic top *U–1* of the urine collection bottle.[1] The metal washer *U–2* is then inserted between the ring *K–5* and the inside of the cap (Fig. 7). A piece of filter paper is inserted into the urine filtration funnel *U–3*. The funnel *U–3* is then placed in the urine collection bottle *U–4;* the latter is threaded into the plastic cap *U–1*. The feces are retained on the $^3/_{16}$ in. screen *I*, which is placed within the large funnel *K–1*. The urine passes through the screen and is filtered into the urine collection bottle. The depressions *U–6* prevent the ledge of the funnel *U–5* from sealing the mouth of the bottle and thus provide a vent.

The cage is supported on the rack *R* by means of 2 horizontally placed rods *J–1*. When the cage is positioned in the rack as shown in Figure 7, the rods lie in the lateral grooves *A–4*.

Comment.—For the most part, the component pieces of this cage are stampings or spinnings and so can be manufactured more economically than the type shown in Figures 3–5. The ease of assembling and dismantling them facilitates their cleaning and decontamination. The false bottom, the hinge and the door lock can be completely disassembled for cleaning.

The funnel of the cage can be removed without disturbing the rat. The false bottom can likewise be removed for cleaning without removing either the funnel or the rat (i.e., the animal rests in the funnel during this time). If the false bottom should become irreparably contaminated with radioisotopes, it can be replaced at minimal cost.

The cages may be adapted to rats of different sizes, or even to mice, by varying the size of the mesh of the false bottom.

IV. Methods for Quantitative Measurement of Water Intake

A number of animal drinking fountains have been described (2–4, 18, 19, 23, 25), and the mechanics of operation has been discussed (12).

[1] Even-Flow nursing bottles are used.

DETAILS OF CONSTRUCTION

The simplest drinking fountain, shown in Figure 8, *A*, consists of a bottle *A–1* fitted with a 1-hole rubber stopper *A–2* and a bent glass (or brass) tube *A–3*. The bottle is mounted on the outside of the cage with a spring clip (*O*, Fig. 1), and the tip of the glass tube *A–4* projects into the cage. The dimensions of the tube are critical; if the inside diameter is too large, the bottle may empty all at once, whereas if it is too small, the bottle does not vent properly. A brass tube with the following dimensions operates properly (4): outside diameter $^3/_8$ in.; wall thickness $^1/_{32}$ in., and outlet hole constricted to $^3/_{32}$ in.

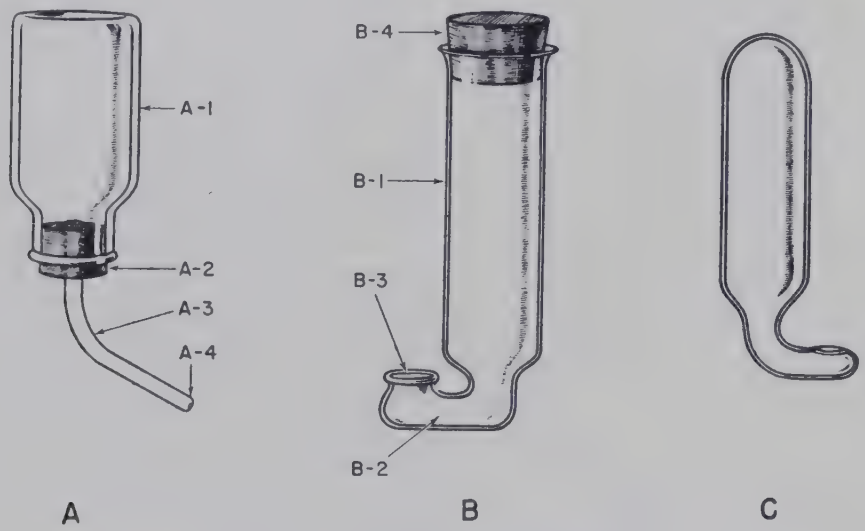


FIG. 8.—Rat drinking fountains.

The glass-blown drinking fountain¶ (Fig. 8, *B*) consists of a cylindrical vertical tube *B–1* and a horizontal portion *B–2*. Water is available through the opening *B–3*. The upper end of the tube is stoppered with a cork or rubber stopper *B–4*. The bottle fits into a spring clip which is mounted on the outside of the cage, and the open end of the fountain *B–3* projects into the cage.

The animal drinking fountain shown in Figure 8, *C*, is similar in construction except that it is completely made of glass.

The cast aluminum drinking fountain* shown in Figure 9 has been in use in our laboratory for the past 5 years. The aluminum casting *D–1* is made of a silicon aluminum alloy (type 43–S), which meets the specifications for cooking utensils. The casting

¶ Available through most laboratory supply houses and the George H. Wahman Company, Baltimore 2.

* Available from the Micro-Metric Instrument Company, P.O. Box 884, Cleveland 22.

consists of an expanded portion *D–1* ($1^{3}/_{4}$ in. in diameter) which is threaded internally to accommodate the calibrated bottle† (*D–10*). The horizontal portion of the drinking fountain *D–2* measures $^{13}/_{16} \times ^{13}/_{16}$ in. and is $1^{1}/_{2}$ in. long. The hole *D–3* in the tip of this horizontal portion communicates with the core *D–11* and the chamber *D–12*. The top of the hole *D–3* is counterbored to increase the drinking area. The bottle *D–10* is held in the spring clip *D–8* which is in turn fastened to the door of the cage *D–7* by the screw *D–14* and wing nut *D–9*. The open end of the drinking fountain *D–3* projects into the cage, as shown; the groove *D–5* fits over

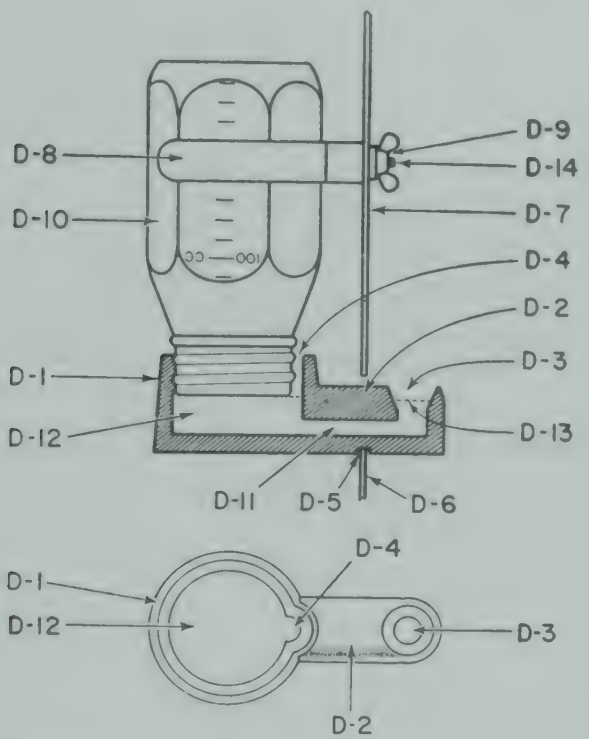


FIG. 9.—Rat drinking fountain: cast aluminum type.

the lower margin of the door opening *D–6*. This assures proper alignment of the drinking fountain.

The water level within the fountain is maintained at a point level with the lowest portion of the bottle *D–13*. A milled groove *D–4* acts as a vent permitting air to enter the drinking fountain. The bottles used are calibrated in cubic centimeters; they are inexpensive and available in a 120 or 240 cc capacity.

To assemble the drinking fountain, the bottle is filled to the top and the aluminum casting *D–1* screwed in place. The opening *D–3* of the fountain is covered with a finger and the bottle inverted

† Even-Flow nursing bottle.

and positioned as shown in Figure 9. A 120 cc calibrated bottle will contain 160 cc when filled to the brim. When the bottle is inverted some of the liquid enters the body of the drinking fountain *D-12*. How much (about 20 cc) is determined by reading the scale and subtracting this correction from subsequent readings. Although the smallest calibration mark is 10 cc, it is fairly easy to estimate the water content to within 2 cc. When more than 120 cc of water is consumed, the water level will be below the lowest calibration mark, and in order to measure the water intake it is necessary to remove the bottle from the clip and invert it.

Comment.—The animal drinking fountain shown in Figure 8, *A*, is inexpensive and has been used in many laboratories. When the dimensions are appropriate, the drinking fountain works well. When mechanical failure occurs, it usually takes place at the most inopportune time. At times the bottle empties completely, leaving the rat without water and thus introducing an error in urine output. Occasionally the liquid does not flow all the way to the tip of the tube, and the animals cannot get the water even though the bottle is full. Water loss may occur when the fur of the rat brushes against the tip of the drinking tube. To measure the volume of water intake, the bottle must be calibrated or the contents of the bottle must be transferred to a graduate cylinder.

The glass-blown drinking fountains (Fig. 8, *B*, and 8, *C*) are less subject to mechanical failure. The fountain shown in Figure 8, *C* is ideal for certain special fluid intake experiments. However, since the drinking tube must be filled through the narrow outlet it is more difficult to fill. A major disadvantage of the glass-blown drinking fountains is the expense. During the course of a single year in which we followed 144 rats, the breakage and replacement cost amounted to almost $100. The bottles were broken during filling, or by the rats pushing them out of the cage.

Most of the commercially available glass-blown drinking fountains have a capacity of 100 cc or less. In using these drinking fountains in our experimental diabetes studies we found it necessary to refill the drinking fountains at least twice a day. Diabetic rats will drink up to 250 cc a day. Refilling these drinking fountains was especially annoying on Sundays. In desperation we devised the calibrated animal drinking fountain shown in Figure 9. These have proved to be most satisfactory. The aluminum castings are moderately priced and the bottles are inexpensive, costing only a few cents. The replacement cost for a year's breakage is small. The bottles are available in two sizes (120 and 240 cc) and so need not be refilled more than once a day. The volume of water consumed can be read directly by using the bottle calibrations. One

disadvantage of this drinking fountain is that after prolonged use an occasional rat will gnaw on the end of the aluminum castings. This will deface the end and necessitates replacement of the fountain.

V. Methods for Quantitative Measurement of Food Intake

The problem of avoiding food spillage is of major importance in metabolic studies. Even when the measurement of food intake itself is not important, food spillage will interfere with the quantita-

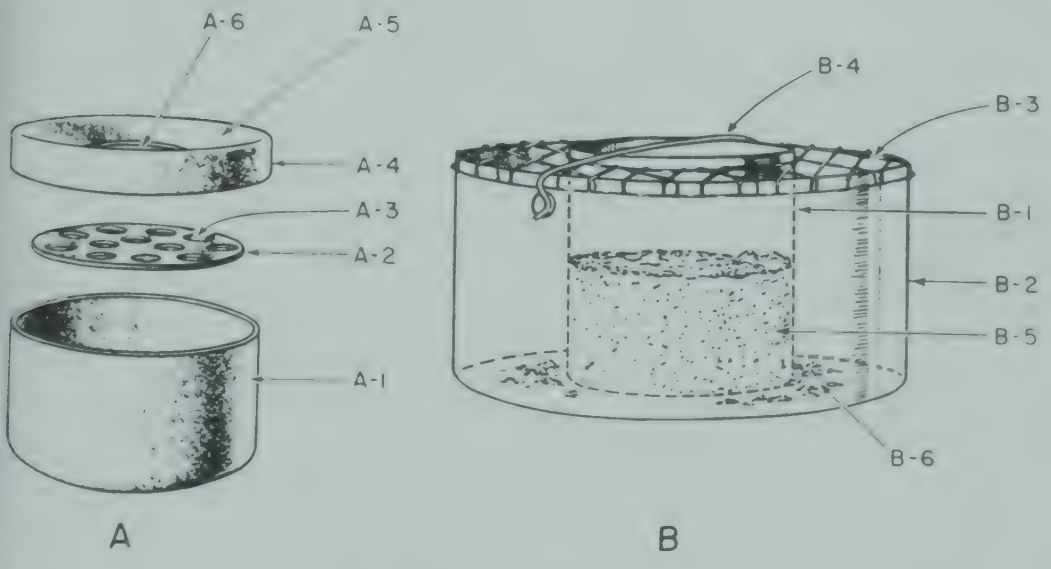


FIG. 10.—Rat feeding devices.

tive collection of urine. Indeed, unless a special non-scatter food container is used, it may be impossible to collect a urine specimen from a normal rat because the spilled food, accumulating in the collection funnel, may soak up nearly all of the urine. The use of pellet food ($^1/_2$ in. pieces) is unsatisfactory when urine collection is required, for even though an external food container is provided, the rats will carry the large food pieces into the cage and drop them through the bottom mesh. The use of a powdered food in an appropriate feeding device has proved much more satisfactory. Feeding by stomach intubation is, of course, the surest quantitative method; it is also the most troublesome.

A number of feeding devices have been described. In some of these, the animal must enter a tubular structure or other restricted area to gain access to the food (1, 5, 18, 19). In this case the entrance to the food cup must be sufficiently restricted so that

the rat cannot turn around and contaminate the food with feces. In those instances where the food cup is placed within the cage, scattering can be minimized by placing a perforated disk over the powdered food (24).

DETAILS OF CONSTRUCTION

The non-scatter feeding dish shown in Figure 10, *A*,‡ is widely used. The powdered food is placed in an aluminum dish *A–1* made of no. 14 gauge spun aluminum. The dish is $3^1/_8$ in. in diameter and

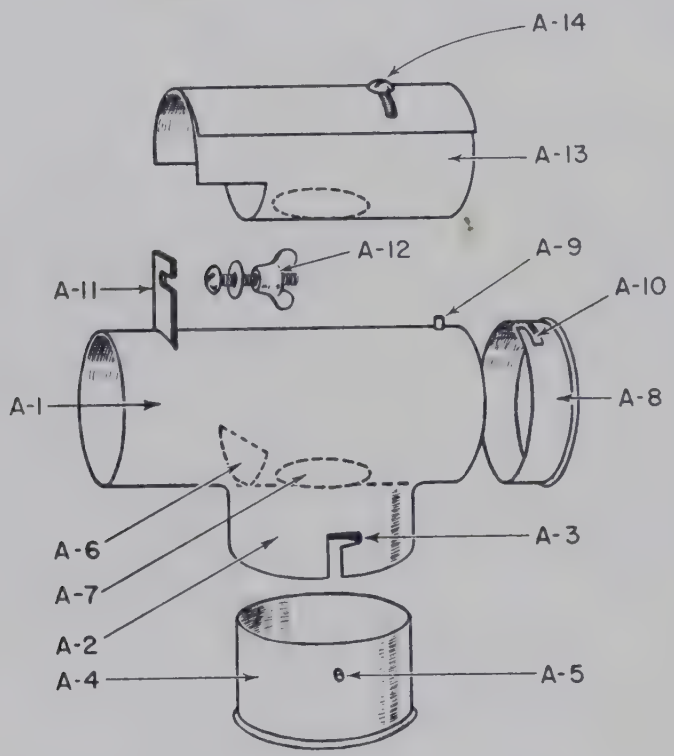


FIG. 11.—Rat feeding device.

$1^5/_8$ in. deep. A perforated disk *A–2* made of Monel metal rests on the surface of the powdered food. The disk is $2^3/_4$ in. in diameter and has a dozen $^9/_{16}$ in. holes *A–3*. This disk keeps the animals from digging into the food and prevents scattering. The cover *A–4* is also made of Monel metal; the top *A–5* is funnel-shaped and narrows to a hole *A–6* that is $1^5/_{16}$ in. in diameter.

A simple non-scatter food cup which has been in use at the University of Rochester is shown in Figure 10, *B*. The food *B–5* is placed in the glass container *B–1* which is in turn placed in the center of a large tin can *B–2* about 4–5 in. in diameter. The metal wire *B–4* serves to retain both the center cup and the wire screen

‡ Available through most laboratory supply houses.

B–3. The screen *B–3* spans the distance between the outside of the food cup and the inside of the tin. Any spilled food *B–6* is collected in the outer container *B–2* and can be weighed with the food cup.

The feeding device (Joy feeding tube (11)) shown in Figure 11 is designed to fit on the outside of the cage, as shown in *N*, Figure 1. It consists of a cylindrical tube *A–1* (Fig. 11) $4^1/_2$ in. long and $1^7/_8$ in. in diameter. The end is closed by the cap *A–8*, locking the peg *A–9* in the groove *A–10*. The food is placed in the cup *A–4* which is $2^1/_8$ in. in diameter and $1^1/_2$ in. deep. This cup slides into the sleeve *A–2* and is fixed by inserting the pin *A–5* into the slot *A–3*. The food cup is easily removed for cleaning and weighing. A hole *A–7* ($1^1/_4$ in. in diameter) in the bottom of the cylinder *A–1* permits access to the food cup. Since rats tend to carry food out of the container, it is useful to have a vertical baffle *A–6* mounted across the inside of the cylinder. For small rats, a liner of adjustable diameter *A–13* is inserted into cylinder *A–1*. The diameter of the tube is increased by loosening the screw *A–14* and expanding the cylinder. The diameter should be sufficiently small to prevent the rat from turning within the cylinder and contaminating the feces.

Comment.—We have used the non-scatter food cups shown in Figure 10, *A*, for the past 4 years with fairly good success. They are inexpensive, easy to clean and easy to maintain. The amount of spillage is small. Occasional rats learn to upset their food dish and do so regularly. In this case the cover must fit tightly and the food dish must be clamped to the inside of the cage. When using these food cups with a high fat diet of a pasty consistency, the rats are unable to get adequate amounts of food through the perforated disk. In this case we have removed the disk *A–2* and occasionally the cover *A–4*.

The feeding dish shown in Figure 10, *B*, might well permit the quantitative recovery of the scattered food, but there would be considerable danger of fecal and urine contamination.

We have had several years' experience with the food cup shown in Figure 11. When pellet food is used, the rats tend to carry the pellets into the cage. When powdered food is used the horizontal baffle *A–6* should be employed, for this markedly diminishes the ability of the rat to carry the powdered food into the cage. The Joy feeding tube, being of more complicated construction than the other feeding devices, is therefore much harder to clean and to decontaminate.

VI. Stomach Intubation

Liquid diets of varying amounts of carbohydrate, fat and protein can be fed by stomach tube (22). Once the technique of

stomach intubation has been mastered and the rats have become accustomed to the procedure, 1 person can intubate more than 40 rats per hour.

TECHNIQUE

The intubation procedure is illustrated in Figure 12. A no. 8

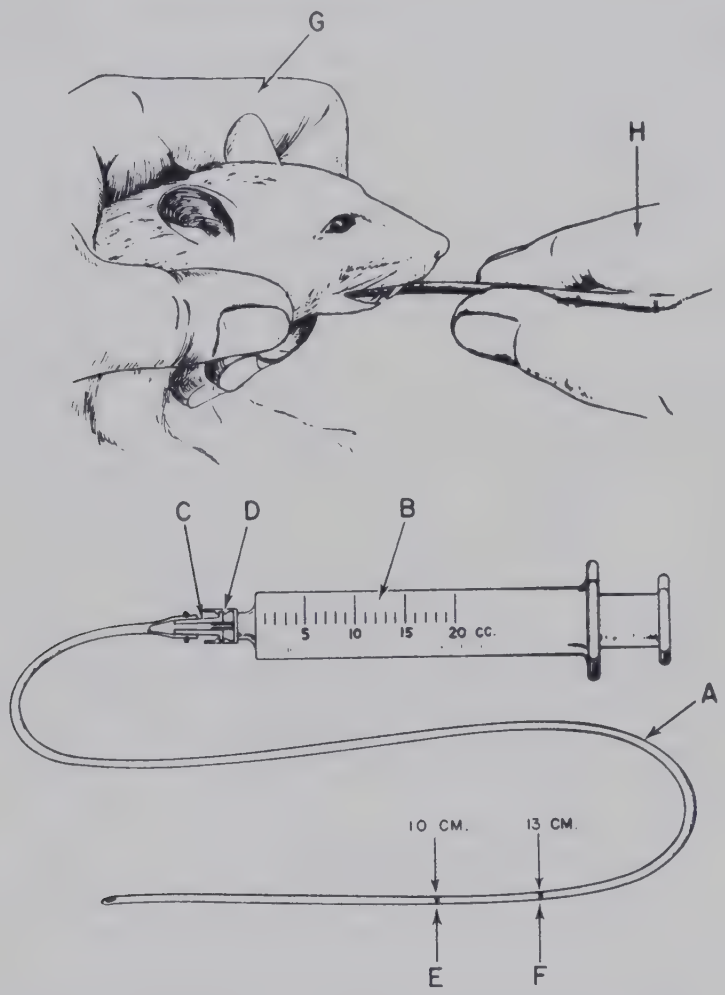


FIG. 12.—Feeding rat by stomach intubation.

French urethral catheter *A* is wired to a standard taper adapter *C*, which is in turn inserted into the Luer-Lok *D* of a 20 cc hypodermic syringe *B*. With India ink the catheter is marked at distances of 10 and 13 cm from the tip (*E* and *F*). When a 300 g rat is intubated to the 10 cm mark, the tip of the catheter is at the cardia of the stomach. When the catheter is inserted to the 13 cm mark, the tip of the catheter lies along the greater curvature of the stomach. (When smaller rats are used, the corresponding distance will, of

course, be smaller.) The rat is placed on the table and the left hand *G* is positioned over the head and the back. The mouth of the rat is forced open by pressing between the mandible and maxilla with the thumb and index finger. The catheter is moistened with water and the tip is placed in the mouth, to one side of the incisor teeth, and rotated between the thumb and index finger of the right hand *H* while it is being inserted. This rotary motion keeps the catheter from bending and facilitates its passage into the stomach. If the rat begins to swallow, it is best to continue advancing the catheter, as this swallowing motion guides the tube into the pharynx. Occasionally the catheter enters the trachea; in this case the animal struggles and the catheter cannot be inserted more than about 7 cm. If the catheter is in the trachea, it should be withdrawn and reinserted. *Caution: Do not begin the injection of food until the catheter has been inserted to a depth of 12–13 cm.*

Comments.—When using a liquid diet it is necessary to feed the rats at least twice a day and to adapt the animals gradually to the forced-feeding procedure. Initially 4 cc of diet may be administered to a 300 g rat; the volume fed is gradually increased by 1 cc per feeding until the full amount (14–16 cc) is administered twice a day (10, 14). Since some of the constituents of the liquid diet are held in suspension, it is necessary to insure adequate mixing of the diet. The diet should be strained through a fine mesh screen to insure its passage through the lumen of the catheter. We have found it convenient to store the amount of diet required for a single set of feedings in an 8-oz nursing bottle. The bottle is thoroughly shaken each time just before refilling the syringe. The syringe is refilled by disconnecting the adapter *C* at the Luer-Lok.

The stomach intubation method is the best for quantitative feeding and for avoiding food contamination of the urine or feces. We have carried 20 rats for over a year on stomach intubation. Since rats do not vomit, there is little doubt that all of the diet administered is consumed. Its primary disadvantage lies in the additional time required for individual feeding of the rats twice a day.

VII. Methods for Quantitative Collection of Urine

A number of devices have been made to facilitate the quantitative collection of urine. Unless an appropriate non-scatter food cup is used, some urine will be lost by absorption on the spilled food. Losses along the sides of the funnel, etc., may be minimized by coating the collection funnel and screen with a hydrophic sub-

stance.§ A number of methods have been devised to separate the urine from the feces. The simplest method (8, 9) is to place a small mesh screen in the urine collection funnel, as shown in Figure 1. More complicated separating devices have been described (1, 6, 7, 16, 18).

When carrying out certain chemical tests on urine it may be desirable to filter the urine before use. This can be accomplished during the collection period. In some instances it is desirable to wash

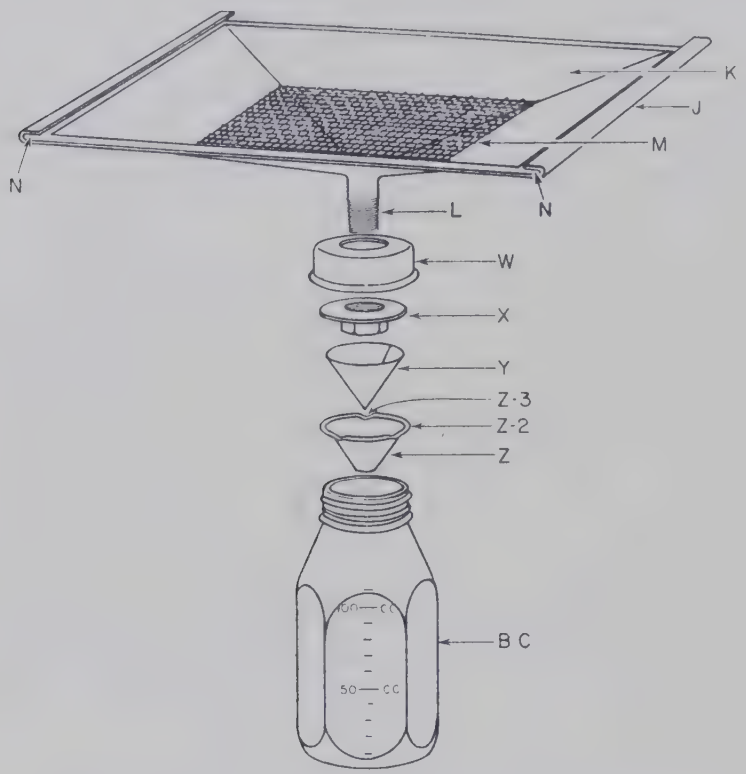


FIG. 13.—Urine collection and filtration system.

down the sides of the funnel to achieve quantitative urine collection.

The urine is usually collected under toluene in order to avoid bacterial decomposition; about 2 cc of toluene is added to each urine collection bottle.

PROCEDURES

In the round wire-mesh-glass funnel metabolism cage unit (Fig. 1), a wire screen *J* ($^1/_4$ in. mesh) is placed within the large funnel B. The feces are retained on the screen whereas the urine passes through the screen into the urine collection bottles *G*. A filter paper cone *K* is placed within the funnel.

§ Paraffin or General Electric Dri-Film, SC-87, may be used.

In the suspended type of metabolism cage (Fig. 3), the feces are retained on the stainless steel screen (*M*, Fig. 13). The plastic cap *W* is attached to the large funnel *K* by screwing the nut *X* over the nipple *L*. A small filtration funnel *Z* is placed within the urine collection bottle *BC* (Fig. 13). The maximum inside diameter of this small funnel is $1^1/_4$ in.; its depth is $^{15}/_{16}$ in. The top of the funnel *Z-2* is flanged so that it rests on the top of the bottle. Three depressions *Z-3* serve to lift the rim of the funnel *Z-2* above the neck of the urine collection bottle and provide an air vent.

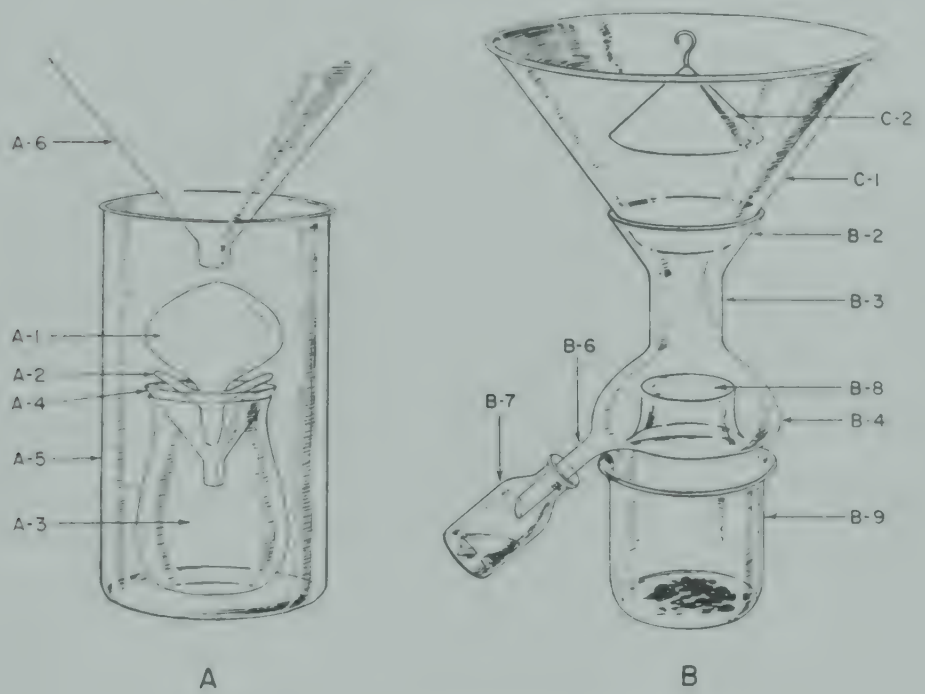


FIG. 14.—Urine collecting systems.

A piece of filter paper 6.5 cm in diameter is folded to form a 60° cone *Y* and placed within the funnel *Z*. The urine collection bottle *BC* with its contained funnel *Z* and filter paper *Y* is then screwed into the cap *W* (as shown in Fig. 4.) By using a nursing bottle which is calibrated in cubic centimeters, the urine volume can be read directly; the volume can be estimated to within 2 cc. When the urine volume is small, it may be necessary to transfer the urine to a graduated cylinder in order to make more precise volume measurement.

The glass sphere method used in separating urine and feces (1, 18) is illustrated in Figure 14, *A*. A large glass globe *A-15* which is $2^1/_4$ in. in diameter, is supported above the neck of the urine collecting vessel *A-3* by means of 3 glass supports *A-2* which rest in the small funnel *A-4*. The diameter of this funnel *A-4* should be

less than the diameter of the globe *A–1*. As the urine drops on the globe, it trickles over the surface and into the collecting vessel *A–3*. The feces and food particles, on the other hand, strike the surface of the globe and are deflected into the outer vessel *A–5*.

The glass trap urine collection device is shown in Figure 14, *B* (6, 13). The trap has a funnel-shaped top *B–2*, a neck *B–3*, and an expanded body *B–4*. The urine is deflected by the baffle *C–2* onto the walls of the large funnel *C–1*, along the walls of the trap *B–2*, into the expanded portion *B–4*, and finally out of the tube *B–6* and into the collection bottle *B–7*. The feces which strike the walls of the large funnel *C–1* are directed toward the center hole of the trap *B–8* and into the container *B–9*.

General comment on urine collection methods.—We have used the screen method (Fig. 13) for separating urine and feces for many years with fair success. If the consistency of the feces is normal, and if the screen fits the funnel well, the separation is adequate. Diarrhea will obviously result in fecal contamination of the urine. Filtering of the urine in situ is advantageous because it saves time. Filtration removes the small food particles or fecal particles.

The urine collection system which is used with the suspended metabolism cage (Figs. 4 and 7) has the advantage of compactness and stability. The rack containing the urine bottles (Fig. 5) can be moved about with little danger of spilling. The urine is collected in a closed system. This minimizes the odor and prevents evaporation. Slight loss of urine will take place in the filter paper. For very precise work it may be advisable to wash the sides of the funnel and the filter paper with water.

The glass sphere and the glass trap methods for separating urine and feces (Fig. 14) would appear to be cumbersome to use. The glass parts are expensive and undoubtedly easily broken. Although the writer has not had any personal experience with these 2 methods of urine collection, they would appear to offer little advantage over the screen-filter-paper method, which he commonly employs.

VIII. Stockade Method for Separation of Urine and Feces

When meticulous separation of urine and feces is required, the stockade cage (20) shown in Figure 15 can be used for male rats. Meticulous separation of urine and feces is necessary in metabolic studies when the fecal excretion of a given substance (such as calcium) is many times greater than the corresponding urinary excretion of this substance. In this case minute fecal contamination in the urine would cause erroneous results.

DETAILS OF CONSTRUCTION

The animal (Fig. 15, *B*) is immobilized in the stockade between 2 parallel lucite plates *A-1*, *A-2*, which are held together by the rods and nuts *A–3*, *A–4*. The chamber *F–1* used for collecting the feces is attached to the rod *F-2* by means of a clamp *F-4*. The rod *F–2* is strapped to the rat with adhesive tape bands around the abdomen and the tail *F–3*, *F-5*. The testes of the rat are placed on a shelf *F–6*. The back of the feces collection receptacle *F–1* is sealed with plastic tape. The urine is collected in a test tube *U–1*, the mouth of which is pressed up against the lower abdomen of the rat and surrounding the penis. The test tube is held in place within

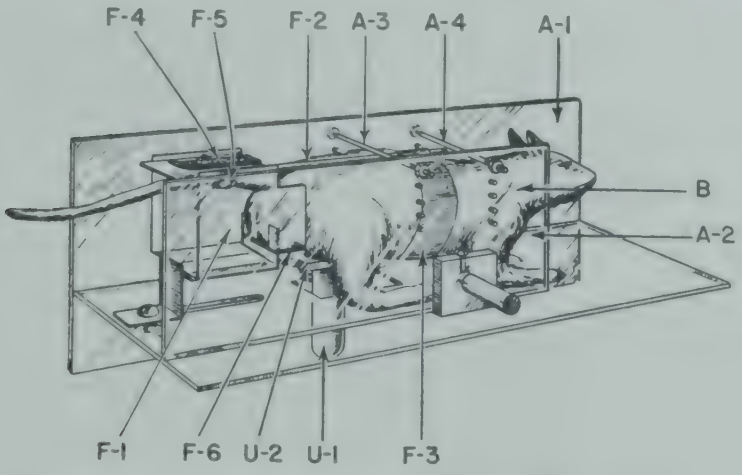


FIG. 15.—Stockade type of metabolism cage.

a plastic plate *U–2* which is strapped to the bottom of the box *F–1* with plastic tape.

Comment.—The stockade cage would seem to "fit the bill" when absolute separation of urine and feces is required. However, this cage is not recommended for long-term experiments; after several days the animals begin to lose weight and they do not eat well (20).

IX. Methods for Quantitative Collection of Feces

When animals are in good health and on an appropriate diet, the feces are well formed. In the screen collection method (Figs. 1, 3 and 6) the feces are removed from the $^1/_4$ in. screen which is placed within the large funnel. When very large rats (500–600 g) are used, the size of the feces may be considerably greater than the mesh which is commonly used for the false bottoms of the cage. Nevertheless the feces are usually forced through the screen when the rat moves about the cage. When large rats are used it is par-

ticularly advisable to make sure that all of the feces have been removed from the false bottom of the cage.

X. Methods for Quantitative Collection of Expired CO_2

The quantitative collection of respiratory CO_2 is essential in many metabolic experiments using isotopic carbon. The collection

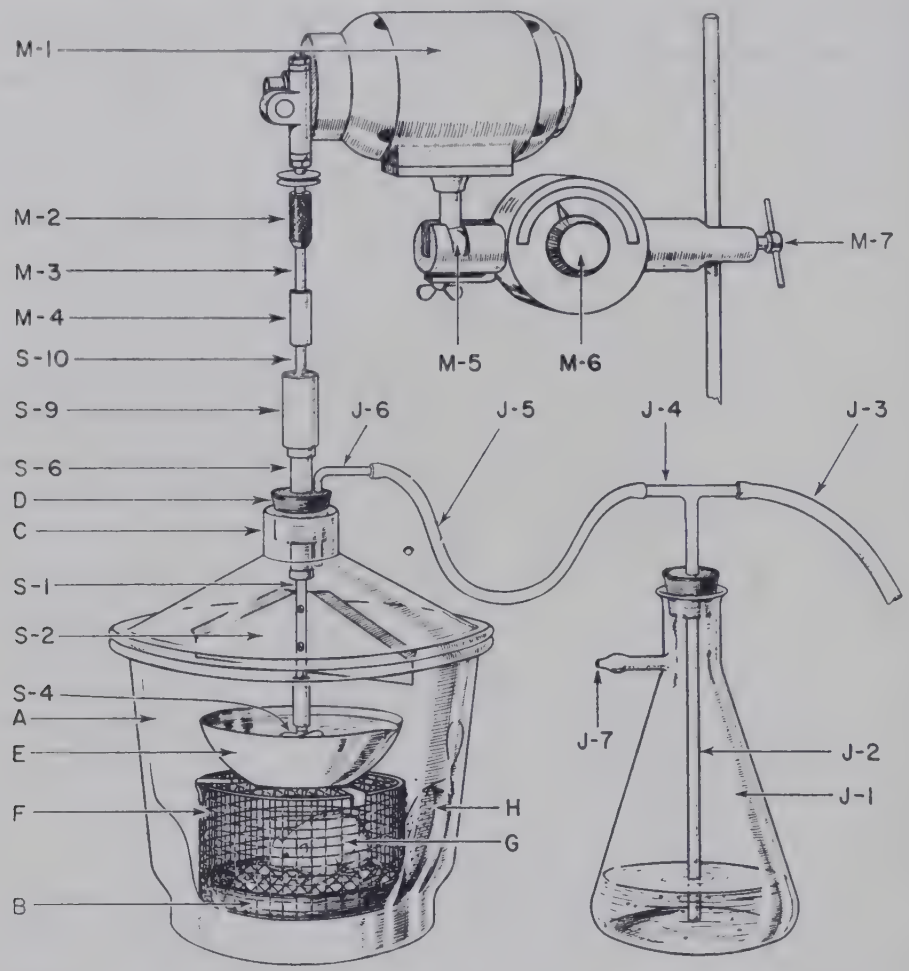


FIG. 16.—Rat metabolism unit used in collecting radioactive CO_2.

unit illustrated in Figure 16 was designed by Drs. N. Lifson and V. Lorber of the University of Minnesota (17) and modified slightly by Dr. W. Sakami of Western Reserve University. This apparatus has been in general use at Minnesota and in the Department of Biochemistry of Western Reserve University for many years.

DETAILS OF CONSTRUCTION

The unit consists of a glass desiccator‖ (*A*) 20 cm in diameter.

‖ Corning no. 3100.

A rat *G* weighing approximately 250 g is placed in a stainless steel wire mesh cage *F* within the desiccator. Two air dryers[¶] *H* are placed in the desiccator. A plastic tray *B* is placed underneath the cage and used to collect the urine and feces. If it is necessary to separate urine and feces, an appropriate screen may be placed within the tray *B*. A 15 cm evaporating dish *E* containing 150 cc of 2N NaOH (CO_2-free) is placed above the cage. A 2-holed rubber

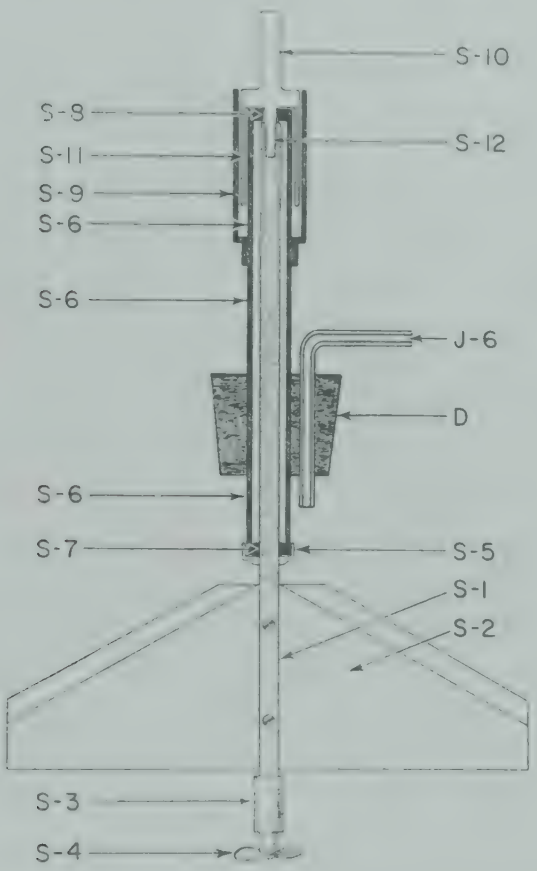


FIG. 17.—Details of construction of stirring apparatus.

stopper *D* is inserted in the opening in the desiccator lid *C*. The air within the desiccator chamber is stirred by the metal fan *S–2*. The alkali in the evaporating dish is stirred by the glass paddle *S–4*. Both of these are driven by the drive shaft *S–1*. The variable speed motor* *M–1* is attached to the drive shaft *S–1* via the chuck *M–2*, the connecting rod *M–3*, the rubber coupling *M–4*, and accessory drive shaft *S–10*. Details of the stirring mechanism are shown in Figure 17. The large metal blades of the fan *S–2* are at-

¶ Eberbach, no. A-4220. The silica gel ordinarily supplied with these air dryers is replaced by anhydrone.

* Type NS-3—1/50 hp, 115 v, 5,000 rpm, Krebs Electrical Manufacturing Co., New York.

tached to the drive shaft *S–1*. The drive shaft housing *S–6*, which contains the ball bearings *S–7*, *S–8*, is inserted through 1 hole of the rubber stopper *D*. The secondary drive shaft *S–10* passes through the upper ball bearing *S–8* and screws into the hole *S–12*, drilled in the primary drive shaft *S–1*. The mercury trap housing *S–9* is pressed on the drive shaft housing *S–6*. Both the mercury trap housing and the drive shaft housing are made of stainless steel. The trough formed between these 2 housings is filled with mercury. The cylindrical portion *S–11* of the accessory drive shaft *S–10* dips into this mercury trough and thus provides an air-tight seal for the stirring apparatus.

When the rat is placed in the cage and the apparatus is assembled and sealed, as shown in Figure 16, the desiccator chamber contains air. The CO_2 which is evolved is absorbed by the NaOH in the evaporating dish, and as the rat utilizes the oxygen, the oxygen is replaced via the tubes *J–3*, *J–4*, *J–5*, and *J–6*. A slightly positive pressure (2–3 cm of water) is maintained in the system by bubbling oxygen through the tube *J–3;* the excess escapes via the tubes *J–2* and *J–7*. A CO_2 content of approximately 0.1% is maintained in the chamber, which introduces a small error in the determination of total CO_2 evolved by the animal.

Comment by Robert Gaunt

Dr. Lazarow is to be commended for having gathered together the assorted literature concerning designs of equipment for metabolic study in rats. Although no one type of equipment can solve the variety of problems with which various investigators are confronted, his discussion can be consulted profitably by anyone trying to build such equipment. It might be pointed out that for acute experiments (e.g., testing of antidiuretic drugs) in which feeding is not a problem, much simpler equipment can be used than that necessary for chronic experiments. An efficient design of both cages and racks for such use is available through the Norwich Wire Works, Inc., Norwich, N. Y.

An inexpensive $6^1/_2$ in. polyethylene funnel is now available which, because it withstands rough handling, can be used to great advantage with small metabolism cages. Dr. Lazarow recommends a urethral catheter for stomach-tubing rats. Perhaps such tubes are essential for forced feeding over long periods, but for the daily intragastric administration of drugs, for instance, we have found a metal tube (hypodermic needle with "ball point" tip) to be completely satisfactory and just about the ultimate in simplicity and durability.

Comment by Dwight J. Ingle

The investigator who aims at meticulous work and efficiency in metabolic studies on the rat will find this excellent contribution to experimental medicine most helpful.

We use a room temperature of 74–78 F rather than 68–72 F indicated by Dr. Lazarow. The important point is that temperature be kept as constant as possible. We have kept healthy rats in metabolism cages at 33 F for months without the development of respiratory disease. As expected, food intake rises as temperature is lowered. Caloric balance is determined in part by the temperature of the surrounding air and by the voluntary activity of the animal. Both can be influenced by the type of cage used. When the cage is very small, activity is restricted. Shielding of the top or sides of a cage can reduce the flow of air to the body of the rat.

We, too, prefer cages and funnels of stainless metal. Absence of corrosion and breakage justifies the expense. When it is desirable to conserve lateral space, a cage of 4 or 5 in. lateral diameter will accommodate most rats. We routinely use a stainless metal cage having the inner dimensions of 4×10×4 in. A cotton plug, rather than filter paper, can be placed in the funnel to filter the urine. Since there is some evaporation of urine from the screen and funnel, we routinely wash the cage at 24 hr intervals. The washings dilute the urine, which is made up to a constant volume of 100, 200 or 250 cc. depending on the size of the rat and the nature of the experiment. The washing of cages can be expedited by use of a plastic garden hose having a spray nozzle. The hose is attached to an extra outlet at the side of the sink. The use of deionized or distilled water may be required for special studies.

The technique of forced feeding of rats by stomach tube is used routinely in all of our balance studies. We consider the extra effort to be well spent. The diet is poured into plastic cups and kept in a freezer until it is thawed just before use. In our experience it is not necessary for the tube to enter the stomach of the rat. Passage through 2 or 3 cm of the esophagus is sufficient. The depressed eye in the tip of the urethral catheter should be clipped away so that a bevel-shaped tip remains which is then smoothed by emery paper or heat. Many investigators have been discouraged from using the technique of forced feeding because of difficulties encountered before skill is acquired. Gentle handling of the rat is imperative. A skilled technician can train a novice to use the technique successfully within a period of a few days.

REFERENCES

1. Ackroyd, H., and Hopkins, F. G.: Feeding experiments with deficiencies in the amino acid supply: Arginine and histidine as possible precursors of purines, Biochem. J. 10: 551, 1916.
2. Blair, A. W., and Carmichael, E. B.: Cage for mice and rats, J. Lab. & Clin. Med. 19: 1020, 1934.
3. Enzmann, E.: A practical type of mouse cage, Science 76: 496, 1932.
4. Farris, E. J. (ed.): *The Rat as an Experimental Animal* (New York: John Wiley & Sons, Inc., 1950).
5. Guest, G. M., Brodksy, W. A., and Nelson, N.: Metabolism cage for rats, with feeding device that minimizes food scattering, Metabolism 1:89, 1952.
6. Gross, L., and Connell, S. J. B.: Separation of excreta from rats, J. Physiol. 57: 60, 1923.

7. Harned, B. K.; Cunningham, R. W., and Gill, E. R.: A metabolism cage for small animals, Science 109: 489, 1949.
8. Hatai, S.: The excretion of nitrogen by the white rat as affected by age and body weight, Am. J. Physiol. 14: 120, 1905.
9. Henriques, V., and Hansen, C.: Ueber Eiseisssynthese im tierkorper: Z, Physiol. Chem. 43: 418, 1904.
10. Ingle, D. J.: The production of alimentary glycosuria by forced feeding in the rat, endocrinology 39: 43, 1946.
11. Reference deleted by author.
12. Lane-Petter, W.: Mechanics of the animal water bottle, Nature, London, 169: 465, 1952.
13. Langham, W. H.: Metabolism of plutonium in the rat, US–AECD–1914, 1946.
14. Lazarow, A.: Glutathione potentiation of cortisone-induced glycosuria in the rat, Proc. Soc. Exper. Biol. & Med. 74: 702, 1950.
15. Lazarow, A.: To be published.
16. Levine, H., and Smith, A. H.: A cage device for the study of ketosis and nitrogen metabolism in small animals, J. Lab. & Clin. Med. 11: 168, 1925–26.
17. Lifson, N., and Lorber, V.: Unpublished.
18. Macallum, A. B.: The relation of vitamins to the growth of young animals, Tr. Roy. Canad. Inst., Toronto 12-13: 175, 1920–21.
19. Owen, S. E.: Small animal metabolism cage, J. Lab. & Clin. Med. 19: 1135, 1934.
20. Peacock, A. C., and Harris, R. S.: Plastic house for the quantitative separation of urine and feces excreted by male rats, Arch. Biochem. 27: 198, 1950.
21. Rapp, K. E.; Skinner, J. T., and McHargue, J. S.: A new type of glass cage for metabolism, J. Lab. & Clin. Med. 31: 598, 1946.
22. Reinecke, R. M.; Ball, H. A., and Samuels, L. T.: High fat and high carbohydrate diets that can be fed to rats by stomach tube, Proc. Soc. Exper. Biol. & Med. 41: 44, 1939.
23. Richter, C. P., and Eckert, J. F.: Mineral metabolism of adrenalectomized rats studied by the appetite method, Endocrinology 22: 214, 1938.
24. Schafer, E. A.: The effects upon growth and metabolism of the addition of small amounts of ovarian tissue, pituitary, and thyroid to the normal dietary of white rats, Quant. J. Exper. Phys. 5: 203, 1912.
25. Thorpe, D. S.: An improved water bottle for small "caged" animals, Science 93: 460, 1941.

B. MOUSE METABOLISM CAGES

ARNOLD LAZAROW, *Western Reserve University*

I. Glass Metabolism Cage

DETAILS OF CONSTRUCTION

The metabolism cage (1) shown in Figure 1 has been adapted

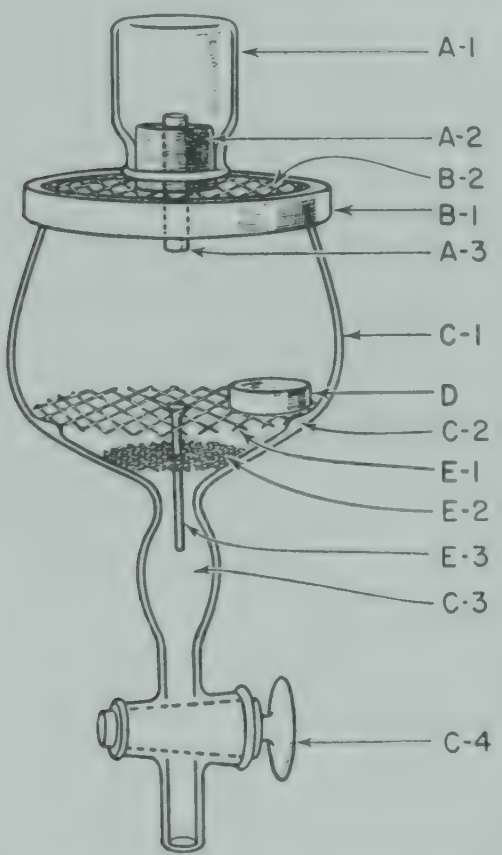


FIG. 1.—Mouse metabolism unit.

from a 500 cc Florence flask. The original flat bottom of the flask is opened and rimmed. This forms the top opening of the cage *B–1*. The neck of the flask is constricted to form the bulb *C–3*. The stopcock *C–4* seals the lower end of the cage. The top of the cage is covered by the screen *B–2*.

The mouse is placed within the glass bulb *C–1* above the screen *E–1*. This screen is supported on 3 glass projections *C–2*. The feces pass through the $^1/_4$ in. mesh screen *E–1* and are retained on the $^1/_{16}$ in. mesh screen *E–2*. Both screens are welded to the brass rod *E–3* which serves as their support. The urine passes through both screens and collects in the bulb *C–3*.

A drinking fountain *A–1* with a 1-hole rubber stopper *A–2* and the drinking tube *A–3* projects through the screen *B–2* and into the cage. The food cup is placed in the metal rim *D* which is soldered to the screen *E–1*. The cage is held in an appropriate rack which supports the bulb *C–1*.

II. Suspended Metabolism Cage

The rat metabolism cage shown in Figures 5 and 6 (pp. 221 and 222) can be adapted to metabolic studies on mice. The false bottom *H* is replaced by a screen which has $^1/_4$ in. holes. The screen *I* (Fig. 6, p. 222) which is placed in the funnel to retain the feces is replaced by one which has $^1/_{16}$ in. holes. Several mice can be put in a single cage of this size. It is possible, of course, to construct a mouse metabolism cage of identical design but of smaller over-all dimensions.

REFERENCE

1. White, F. R.: Sources of tumor proteins: II. Nitrogen balance studies of tumor bearing mice fed a low-nitrogen diet, J. Nat. Cancer Inst. 5: 265, 1945.

C. DOG METABOLISM CAGES

I. Storage and Metabolism Cages

N. R. BREWER, *University of Chicago*

A metabolism cage is one so designed that the amount of urine or feces excreted may be estimated with some degree of accuracy. It is desirable that the cage be so constructed that the urine undergoes a minimum amount of evaporation, no chemical change, no dilution with drinking water and a minimum of contamination with extraneous material.

Design and construction should be such that there will be few corners, no dirt traps and no hiding places for vermin. To effect this ideal, welded construction should be used, filling and grinding the joined parts so that the union is smooth and continuous. Wherever supporting corners are used, no crevices should exist.

Stainless steel is the material of choice for fabricating dog metabolism cages; it does not react chemically with urine, cleaning is made easier and the cost of cleaning reduced, there is no rusting out and upkeep costs are negligible. The type of stainless steel used should contain 18% chrome and 8% nickel (type 18-8 302). A satin finish (2B) is satisfactory.

A fabricator of dog metabolism cages must consider that the design of some experiments calls for long periods of confinement, during which the dogs leave the cages for brief periods only. Too, some researchers insist that the only type of cage for dogs should be metabolism cages, reasoning that space can thus be conserved.

Many researchers tend to place 1 cage on the top of another, reducing the amount of space required to house dogs. The combination invariably demands changes in design. The result is uniformly a construction that is more costly to clean and service. In one study made at the University of Chicago, it cost $7,000 more a year just to clean 100 double-deck metabolism cages than it did to clean 100 double-deck storage cages. That study convinced us that storage cages should be used for storing dogs, and that metabolism cages should be used only when indicated.

DETAILS OF CONSTRUCTION

The simplest type of dog metabolism cage is essentially a large square funnel with built-up walls and a cover (Fig. 1). The dog stands on a false bottom. By means of a long-handled brush, or a

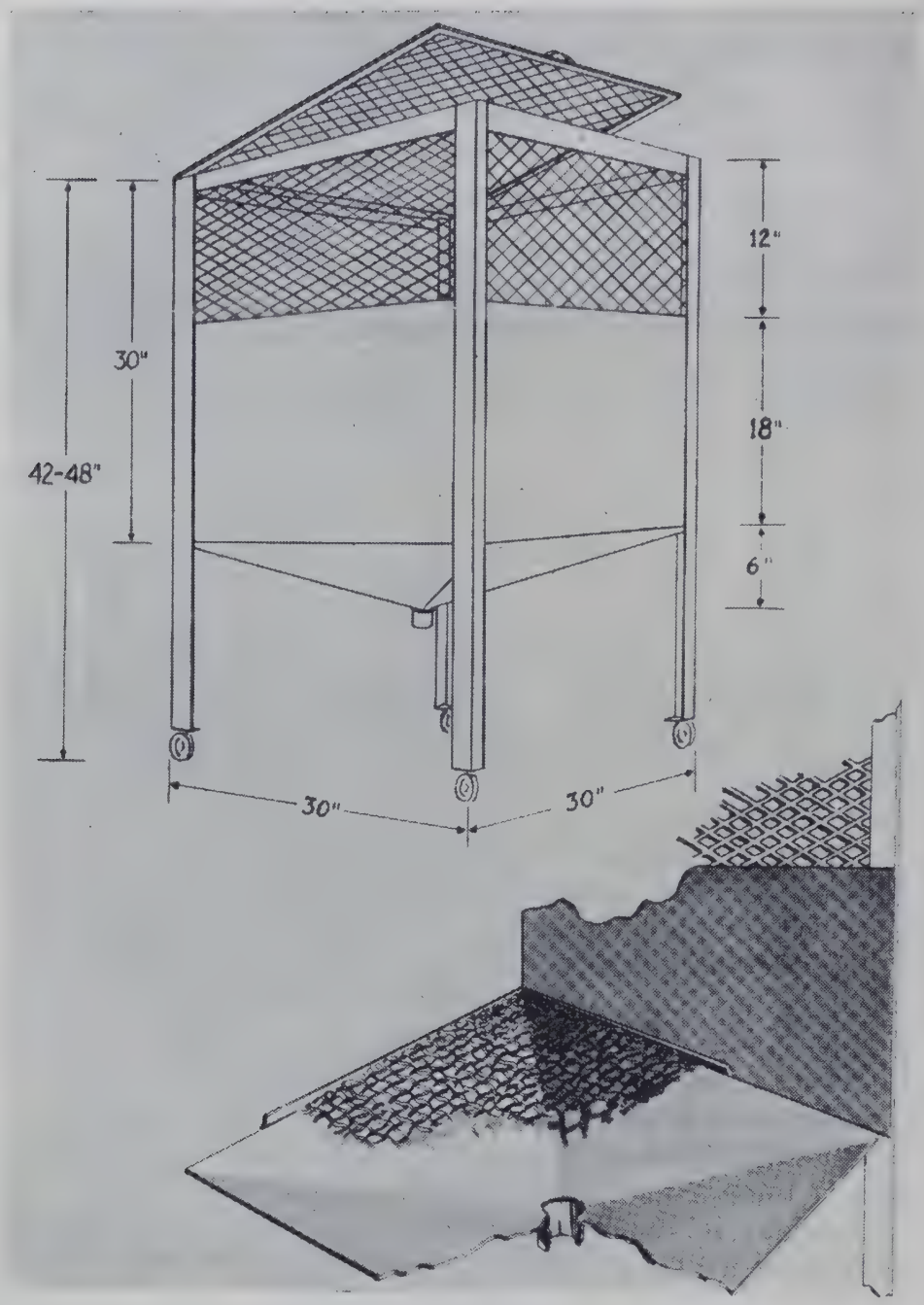


FIG. 1.—Dog metabolism cage.

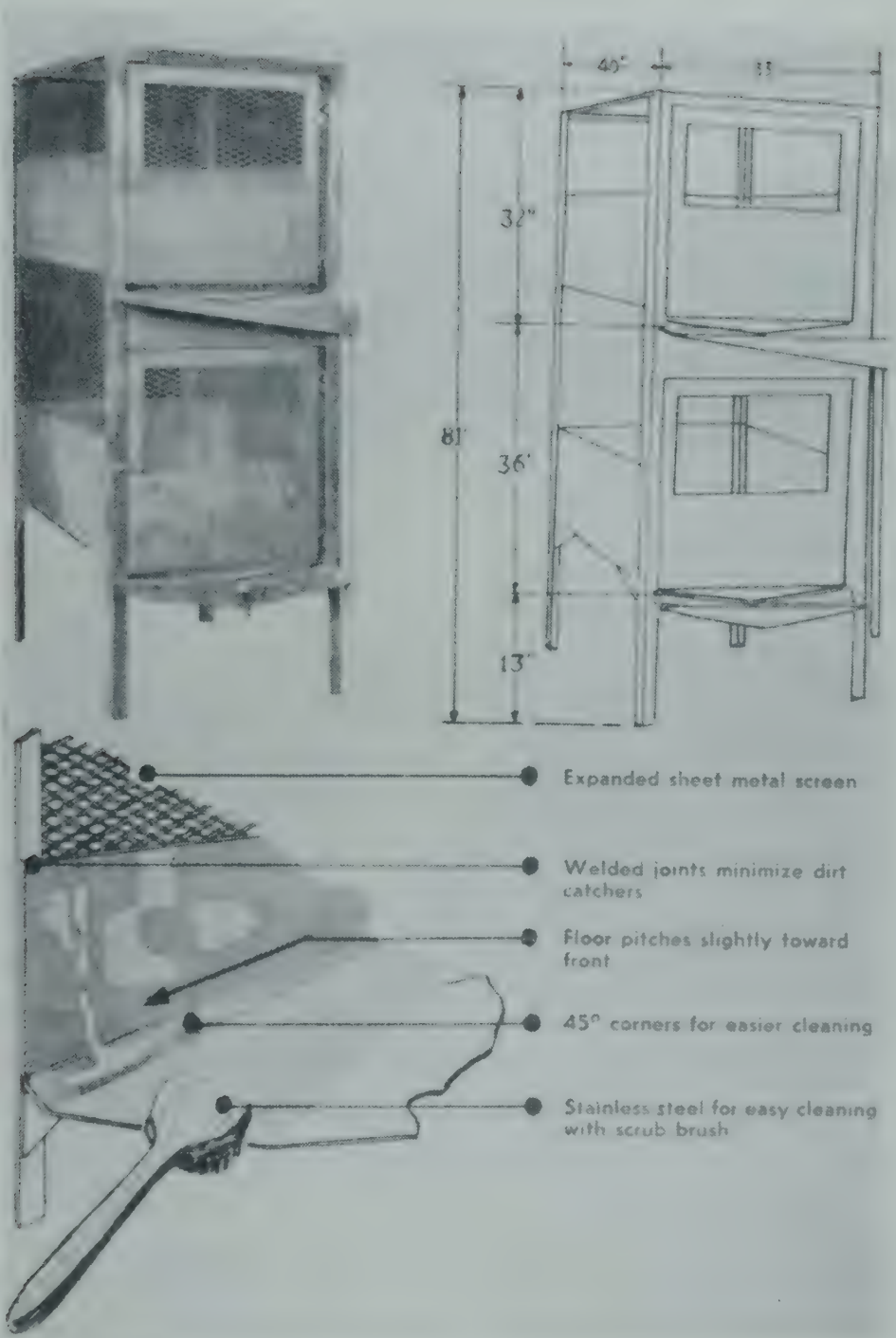


FIG. 2.—Combination storage and metabolism cage.

fountain brush, such a cage is relatively easy to clean. This simple type has most nearly met all of the needs specified for the ideal metabolism cage.

The popular demand for double-deck metabolism cages offered a challenge to improve the cumbersome double-deckers in existence

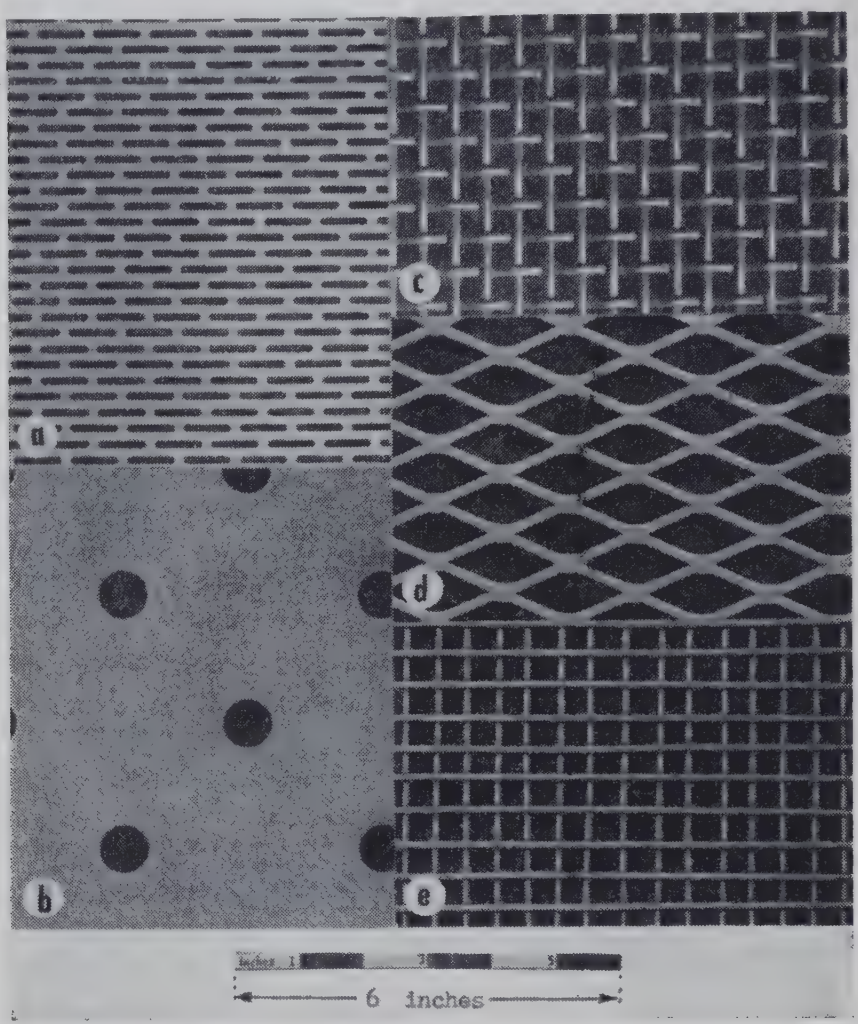


FIG. 3.—Types of false bottoms for dog cages. *a* and *b*, perforated metal type; *c*, woven wire type; *d*, rolled expanded metal type; *e*, welded wire type.

up to 1945. The cage that has met most of the critical needs of researchers in this respect is that pictured in Figure 2. It has been in use at the University of Chicago and at Michael Reese Hospital, Chicago, for 4 years. The cage is 33 in. wide, 40 in. deep and 81 in. high.

The bottom and solid portion of the sides of the cage are bent from 1 piece. The floor pitches forward to the front trough for $1^1/_4$ in. The perpendicular portion of the sides of the piece measures 15 in. in back and $16^1/_4$ in. in front. There is then a bend of 45° for

a perpendicular distance of 2 in. This is followed by a bend to form the floor of the cage, sloping toward a center crease that is $^3/_8$ in. lower than the floor at the sides.

The collecting trough for the top cage tapers from a point level with the floor of the cage at 1 end to a depth of 4 in. at the collecting end. The collecting tube, $1^1/_2$ in. stainless steel, carries the urine to the specimen bottle near the floor. The collecting trough for the bottom cage tapers to a level 2 in. lower toward the center of the cage than at the ends, where another $1^1/_2$ in. collecting tube is placed.

False bottoms have been a special problem in all animal work and types are under continuous study at several institutions. We are acquainted with expanded metal false bottoms, perforated metal false bottoms (with both square and round perforations) and woven wire false bottoms. The possibilities for variations in types (Fig. 3) are many, particularly in the perforated metal types. In our hands a woven wire false bottom, using a 9 gauge, 2 × 2 mesh, has been satisfactory. We are also experimenting with a 9 gauge welded wire false bottom and with several types of perforated metal false bottoms.

II. Circular Metabolism Cage Suitable for Radioisotope Balance Studies

ARNOLD LAZAROW, *Western Reserve University*

For the most part, the rectangular dog metabolism cages now used are modifications of the cage design introduced almost 50 years ago (2, 1). The circular cage described by Hansard (3) (Fig. 1) is the first radical change in design in recent years.

The cage consists of 2 concentric wire cylinders *A* and *B*. The dog is placed between these cylinders on a specially designed false bottom which consists of 2 concentric plywood hoops.

DETAILS OF CONSTRUCTION

The outer cylinder *A*, 29 in. in diameter and 24 in. high, is made by rolling a sheet of no. 9–11 gauge corrugated metal into a cylinder and bolting the edges together. The inner cylinder *B* is made of $^1/_2$ in. mesh woven wire.

The false bottom is fabricated from two concentric plywood hoops $^1/_2$ in. thick. The outer hoop *C–1* is 24 in. in outside diameter and 3 in. wide, whereas the inner hoop *C–2* is $13^1/_2$ in. in outside diameter and 3 in. wide. Clearance between the inner and outer hoops is $2^1/_4$ in., and there is a $6^1/_2$ in. hole in the inner hoop *C–2* to accommodate cylinder B. The plywood hoops are held together by 3 metal rods *C–4*. The rods are $^5/_{16}$ in. in diameter and are bolted to the undersurface of the wooden hoops. The outer hoop *C–1* is attached to the outer cylinder of the cage *A* by 3 supporting rods *C–4;* 1 of these rods (not illustrated) may be retracted into the hoop and thus permits the removal of the false bottom.

The frame *D–1* which supports the cage is made of 2×4's. The cage is supported at a convenient height by the angle iron legs *D–2*. The outer cylinder *A* is attached to the frame by 4 L-shaped metal strips *D–3*. The angle iron *H* which is attached to the frame supports the funnel *F* and the feces separator screen *E*.

The food and water cups *G* (only 1 is illustrated in the figure) are placed on the outside of the cage. A flap of the outer cylinder wall *A–2* (6×6 in.) is cut and bent outward. The dog has access to the feeding dish through this opening. The margins are protected by heavy canvas *A–3* sewed around the raw edges.

The top of the cage *L* is made of 1×2 in. wooden strips (see inset) and is attached to the back of the cage by the hinge *K*.

A removable cylindrical metal liner (not illustrated), 23 in. in

diameter and 17 in. high, may be inserted on the inside of the cage to prevent contamination through the outer cylinder.

A strippable paint* may be sprayed on the metal liner, false floor, feces separator *E*, and funnel *F* in order to facilitate decontamination.

Comment.—Dogs weighing between 7 and 12 kg can be used. As

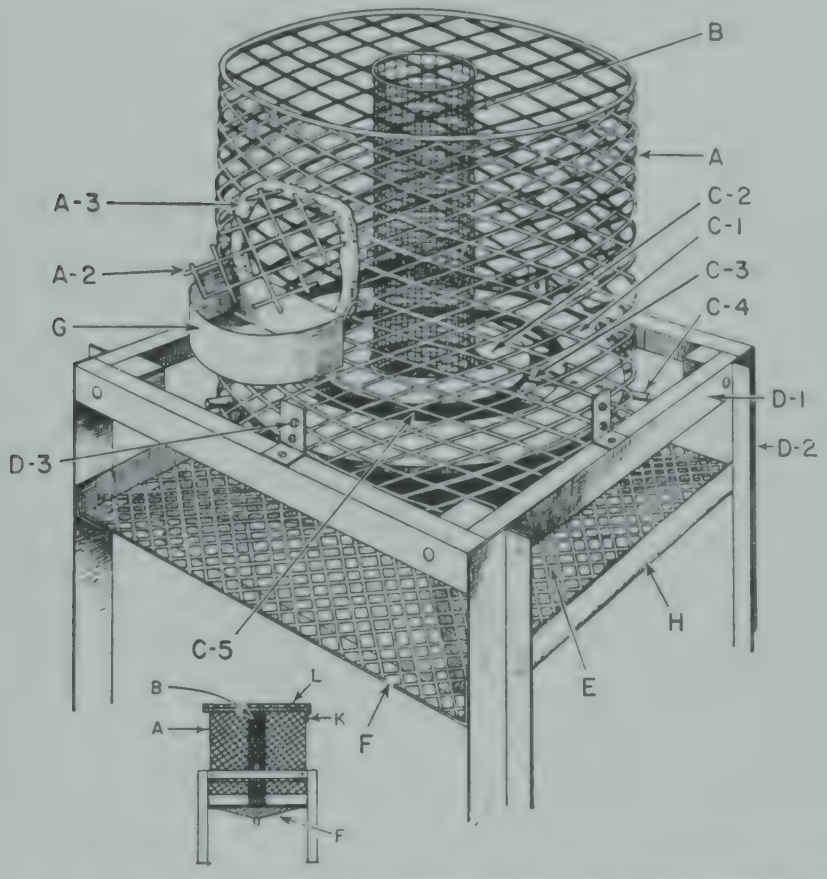


FIG. 1.—Round type of metabolism cage for radioisotope studies.

the animal, standing on the 2 wooden hoops *C-1*, *C-2*, voids, the urine passes through the bottom of the cage, through the screen *E* and into the funnel *F*. The feces pass through the circular opening *C-3* (between the wooden hoops) and collect on the screen *E*. Hansard (4) has used these cages for over a year and reports that smearing of the feces on the wooden hoops has occurred in only 2 or 3 instances. Although one would expect that the dog's feet might slip between the plywood rings, this does not apparently happen (4).

* A strippable paint (Cocoon) is available from the Hollingshead Corporation, Camden, N. J.

Comment by Sam L. Hansard

The drawing of the metabolism unit is excellent. I think it should be stressed that this metabolism unit, although designed for radioisotope balance studies, is equally adaptable for any metabolism or nutritional study with dogs of either sex when quantitative separate collection of urine and feces is desired.

We feel that the circular metabolism unit for dogs is unique for balance studies involving radioisotopes and, when properly adjusted, involves a minimum of urine evaporation and cross-contamination that proves its adaptability for any study requiring quantitative separate collection of urine and feces.

REFERENCES

1. Bliss, A. R., Jr.: A metabolic cage for dogs, J. Am. Pharm. A. 18: 681, 1929.
2. Gies, W. J.: An improved cage for metabolism experiments, Am. J. Physiol. 14: 403, 1905.
3. Hansard, S. L.: A metabolism unit designed for radioisotope balance studies with dogs, Science 117: 301, 1953.
4. Hansard, S. L.: Personal communication, 1953.

D. METABOLISM CAGES FOR MONKEYS*

ARNOLD LAZAROW, *Western Reserve University*

The metabolism unit illustrated in Figures 1–3 is a modification of the cage design used by Dr. C. H. Rammelkamp, of the Department of Preventive Medicine, Western Reserve University.

DETAILS OF CONSTRUCTION

The rectangular cage is 18 in. wide, 30 in. high, and 22 in. from front to back. The frame of the cage *A–7* (Fig. 1) is made of $1^1/_4 \times 1^1/_4 \times {}^1/_8$ in. angle iron, welded at the seams. The top of the cage *B* is made of expanded metal, ${}^{16}/_{18}$ gauge, with ${}^3/_4$ in. openings. The back and sides of the cage *A–1* are made of galvanized sheet metal, no. 20 gauge. The lower portion of the side walls *A–1* is bent as shown in *A–5* and *A–6* of the lower inset, Figure 2. This brings the edge *A–6* away from the angle iron *A–2* and facilitates the flow of urine into the collection pan *F–1*. Three sets of horizontal angle irons, *A–2*, *A–3*, *A–4*, are welded across the base of the side walls of the cage. These support the false bottom of the cage *E–1* and the urine collection funnel *F–1*.

The door of the cage is made of $1 \times 1 \times {}^1/_8$ in. welded angle iron and ${}^{16}/_{18}$ gauge expanded metal (${}^3/_4$ in. openings). The door is fastened by means of the lock *D–8*. An opening *D–3* in the door of the cage which measures $7^1/_2 \times 3^1/_4$ in. permits the insertion of a food and watering dish. The feeding pan *G–1* (Fig. 2) is held in place by the band *D–4* and the lock *D–5*.

The false bottom of the cage *E–1* (Fig. 1) is made of expanded metal (${}^{16}/_{18}$ gauge and ${}^3/_4$ in. openings). The edges *E–2* are reinforced by a ${}^3/_8$ in. metal rod. The false bottom is supported by the angle irons *A–2* and is held in place within the cage by the stops *D–11* and *D–12*. When these stops, which are mounted on rivets, are rotated through 90°, the false bottom and pan may be inserted.

The pan of the cage *F–1* measures $22 \times 17^1/_2 \times 1$ in. and is made of stainless steel. The botton of the pan slopes from the back to front and from the sides to the center. The pan should have a pitch of about 2 in. so that the fluid will gravitate toward the nipple *F–2*. The nipple is made of ${}^1/_2$ in. stainless steel tubing. When the pan is in place (Fig. 2), this nipple *F–2* lies in front of the rack support angle iron *R–2* and over the accessory urine collection receptacle *F–11*.

* Built by the Reister Thesmacher Company, Cleveland.

An accessory stainless steel collecting receptacle *F–3* (or *F–11*, Fig. 2) is mounted on the front supports of the rack *R–2*, *R–3* by the bracket *F–5* (upper inset, Fig. 2). The construction of this

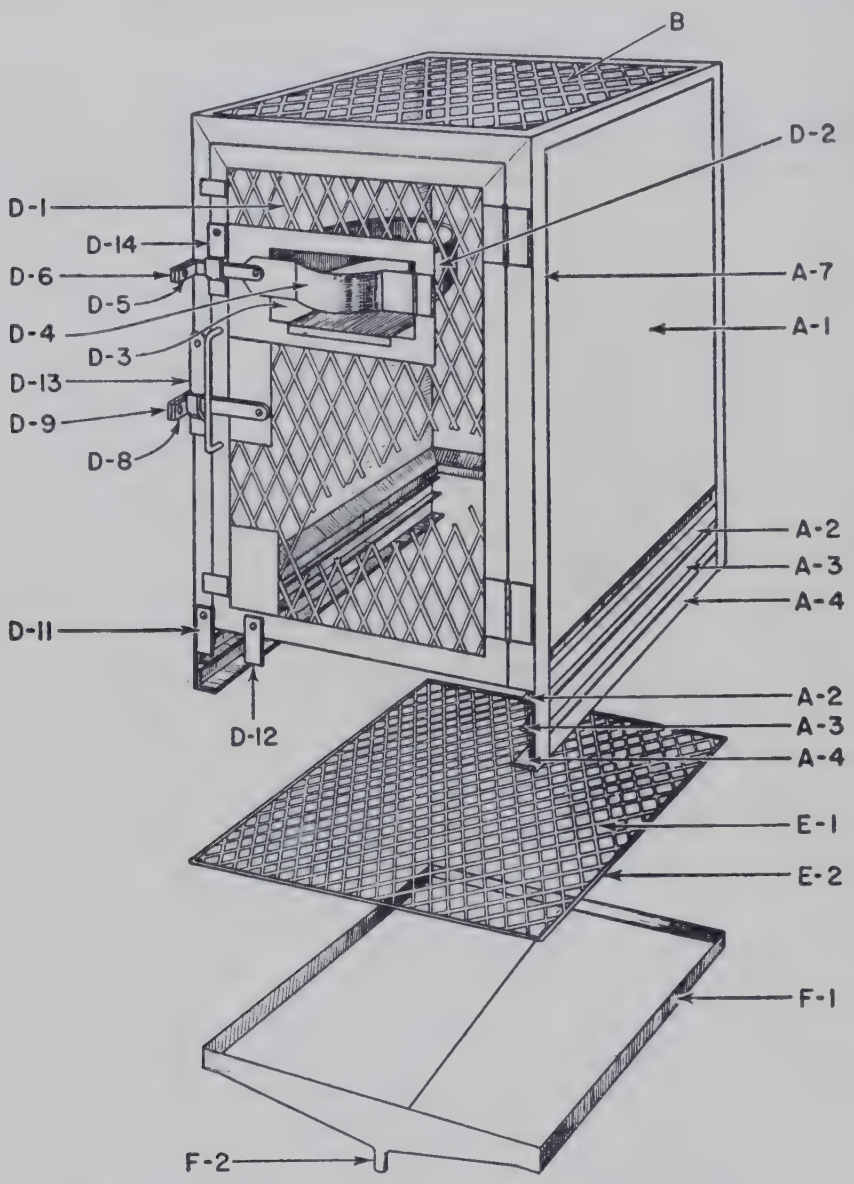


FIG. 1.—Monkey metabolism cage; component parts.

receptacle differs slightly for the upper and lower cages. A $^1/_2$ in. stainless steel tube (*F–6*, or *F–9*) is reamed to fit over the outlet tube *F–4* of the accessory collecting receptacle. The urine, which collects in the pan *F–1* drains into the receptacle *F–3* into the tube *F–6* and into the urine collection bottle *F–7*. The urine collection bottle is placed on the shelf *R–4*.

The door lock is made of a rectangular bar stock ($1^1/_4 \times ^1/_8$ in.).

The cage door is opened by bringing the bar *D–13* to a horizontal position and lifting the bar *D–8* out of the groove *D–10*. Inasmuch as 2 hands are required to open this lock, most monkeys are unable to manipulate it. For those monkeys who have succeeded in open-

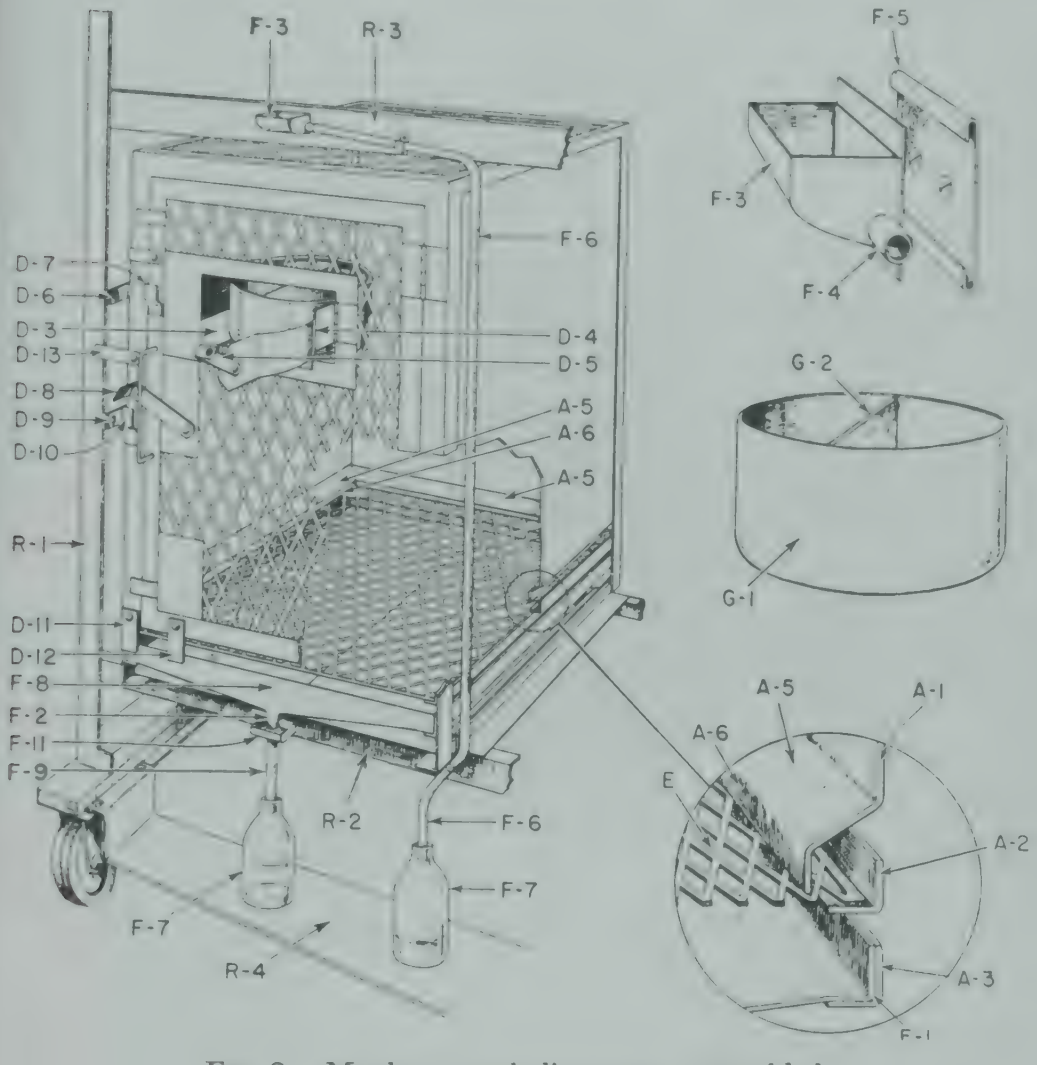


FIG. 2.—Monkey metabolism cage; assembled.

ing this lock, a hasp with a powerful spring is placed through the holes in the bars *D–8* and *D–9*.

The feeding dish (middle inset, Fig. 2) is made of a stainless steel surgical bowl, $7^{5}/_{8}$ in. in diameter and 3 in. high. It is divided into 2 fluid-tight compartments by the plate *G–2*. The food door lock *D–5* is similar in construction to the cage door lock. The pan is inserted into the cage as shown in Figure 2, through the opening *D–3*. The pan rests on a shelf and is retained within the cage by the metal band *D–2* (Fig. 1). The upper edge of this band is rolled to

form a $^1/_4$ in. ledge which fits over the top rim of the pan. This prevents the monkey from lifting the pan out of its retainer.

The rack shown in Figure 3 is made of $1^1/_4 \times 1^1/_4 \times ^1/_4$ in. angle iron. Six metabolism units are assembled. The rack is mounted on 6 in. casters to facilitate movement. The urine collection bottles are assembled on the shelf *R–4*.

Comment.—We have used this monkey metabolism cage for over a year. It has proved fairly satisfactory in most respects. The monkeys are collared and chained. The chain is brought out through the door of the cage and the hasp is attached to the door

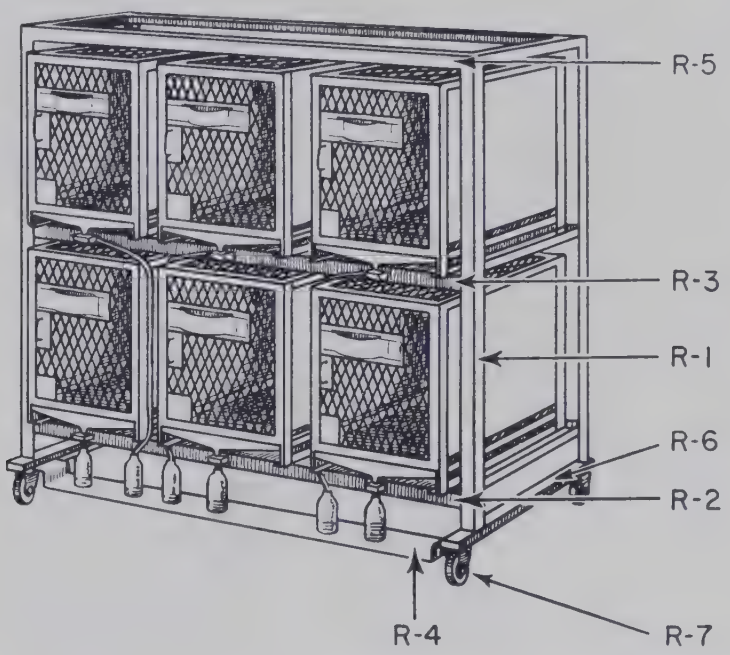


FIG. 3.—Rack for monkey metabolism cages.

handle. When the monkey is to be removed from the cage for blood determination, etc., the animal is pulled against the door of the cage by pulling on the chain. With the monkey held tightly against the door, the cage is opened and the monkey's 2 arms are brought behind his back. With the animal thus immobilized the chain is released and the animal removed.

The monkeys usually urinate while sitting in a squatting position on the false bottom. Usually all of the urine voided is collected in the funnel and there is little danger of urine loss through the perforations in the door. In some instances, it may be advisable to line the lower half of the door with a lucite sheet $^1/_8$ in. thick.

The pan and the false bottom can be removed from the cage for

cleaning without disturbing the monkey. The pan is cleaned first, then the false bottom is removed. The monkey is allowed to stand in the pan while the false bottom is being cleaned. There should be sufficient clearance between the false bottom and the cage door so that the feces will clear the bottom of the door as the false bottom is removed.

The bracket *D-2* (Fig. 1) used to retain the feeding pan is often used by the monkey as a perch. This may result in the loss of urine or in the contamination of the food with urine and feces. It would be preferable to eliminate this pan and use external feeding and watering devices of the type used in rat metabolism studies. This possibility is being explored.

SUBJECT INDEX

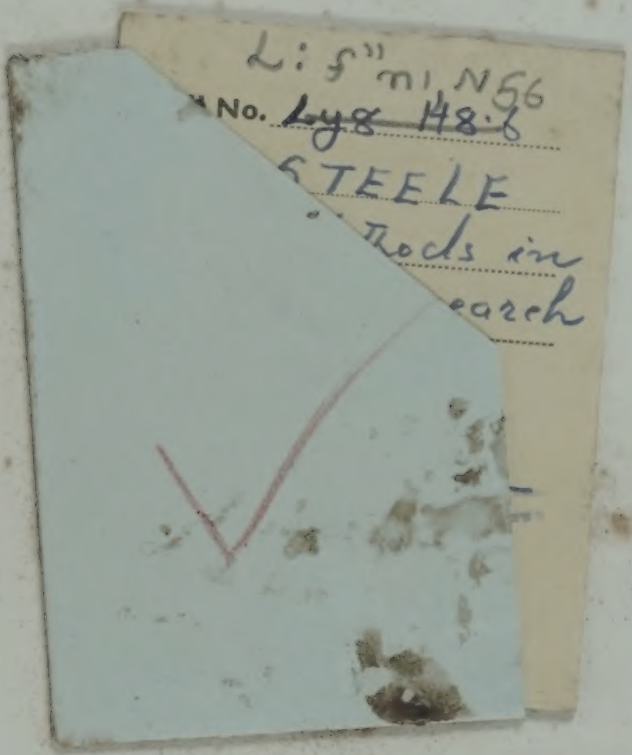
[Page numbers in bold face indicate original contributions to this volume.]

D

T

NAME INDEX

[Page numbers in bold face indicate original contributions to this volume.]


L: 5" m. N56
No. Ly8 48.6
STEELE
rods in
earch